machine learning & knowledge extraction

MDPI

Article

Assessing Interaction Quality in Human–AI Dialogue: An Integrative Review and Multi-Layer Framework for Conversational Agents

Luca Marconi [1,*], Luca Longo [2] and Federico Cabitza [1,3,*]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Via Sarca 336, 20126 Milan, Italy
[2] The Artificial Intelligence and Cognitive Load Research Lab, University College Cork, College Rd, T12 YN60 Cork, Ireland
[3] IRCCS Ospedale Galeazzi-Sant'Ambrogio, Via Cristina Belgioioso 173, 20157 Milan, Italy
* Correspondence: luca.marconi@unimib.it (L.M.); federico.cabitza@unimib.it (F.C.)

Abstract

Conversational agents are transforming digital interactions across various domains, including healthcare, education, and customer service, thanks to advances in large language models (LLMs). As these systems become more autonomous and ubiquitous, understanding what constitutes high-quality interaction from a user perspective is increasingly critical. Despite growing empirical research, the field lacks a unified framework for defining, measuring, and designing user-perceived interaction quality in human–artificial intelligence (AI) dialogue. Here, we present an integrative review of 125 empirical studies published between 2017 and 2025, spanning text-, voice-, and LLM-powered systems. Our synthesis identifies three consistent layers of user judgment: a pragmatic core (usability, task effectiveness, and conversational competence), a social–affective layer (social presence, warmth, and synchronicity), and an accountability and inclusion layer (transparency, accessibility, and fairness). These insights are formalised into a four-layer interpretive framework—Capacity, Alignment, Levers, and Outcomes—operationalised via a Capacity × Alignment matrix that maps distinct success and failure regimes. It also identifies design levers such as anthropomorphism, role framing, and onboarding strategies. The framework consolidates constructs, positions inclusion and accountability as central to quality, and offers actionable guidance for evaluation and design. This research redefines interaction quality as a dialogic construct, shifting the focus from system performance to co-orchestrated, user-centred dialogue quality.

Keywords: interaction quality; human–AI interaction; conversational agents; chatbots; human–AI dialogue; large language models (LLMs); user experience (UX)

Academic Editors: Alex Doboli, K. Wendy Tang and Simona Doboli

Received: 18 November 2025
Revised: 8 January 2026
Accepted: 14 January 2026
Published: 26 January 2026

1. Introduction

Conversational agents are increasingly transforming how users engage with digital systems across diverse domains, including healthcare, education, customer service, and public administration [1–4]. Driven by advances in large language models (LLMs), these systems have evolved from scripted interfaces into more autonomous interlocutors capable of managing open-ended, multimodal, and goal-oriented interactions [5–9]. As these systems gain prominence, defining what constitutes quality in human–artificial intelligence (AI) dialogue has become a central concern in human–computer interaction (HCI), AI ethics,

and usability engineering [10–13]. Traditional evaluation approaches—centred on technical metrics or task success—are increasingly seen as insufficient [14,15]. Instead, interaction quality is now widely recognised as a multidimensional construct shaped not only by pragmatic performance, but also by users' subjective experiences of trust, transparency, social presence, and inclusion [12,15].

Although recent years have seen a proliferation of empirical studies assessing conversational agents, this research remains fragmented. Constructs such as usability, usefulness, satisfaction, trust, and engagement are inconsistently defined, heterogeneously measured, and unevenly integrated across domains and modalities [12,16,17]. Existing reviews typically focus on narrow application areas or specific interaction modalities, offering limited cross-cutting synthesis [1,11,18–20]. Critically, while many studies have examined user perceptions of conversational agents, few have foregrounded these perspectives within a unified framework of interaction quality [21]. As a result, a conceptual gap persists between domain-specific findings and a cohesive understanding of the dimensions that underpin high-quality human–AI conversations.

To address this gap, we conducted a structured, integrative review of 125 peer-reviewed empirical studies, published between 2017 and 2025, on user-perceived interaction quality in human–AI dialogue. Our review spans both text-based and voice-based agents, including LLM-powered systems, and synthesises user-driven evaluative dimensions across diverse contexts. We systematically identify core constructs, measures, and design factors that shape user evaluations, and examine how these elements cluster into consistent patterns of quality judgment. Based on this synthesis, we propose a multi-layered interpretive framework that reconceptualizes interaction quality not as a static system attribute, but as an emergent property of human–AI dialogue. We argue that perceived interaction quality arises from the ongoing co-orchestration between an agent's technical capabilities—its ability to generate relevant, coherent, and timely contributions—and its contextual alignment with user goals, expectations, constraints, and ethical considerations. From this perspective, interaction quality is shaped by dialogic processes such as turn-taking, repair, explanation, and role framing, rather than being reducible to isolated performance metrics.

Our findings reveal three interrelated layers of user judgment: a pragmatic core comprising usability, task effectiveness, and conversational competence; a social–affective layer encompassing social presence, warmth, and synchronicity; and an accountability and inclusion layer defined by explanation efficacy, transparency, accessibility, and fairness. We formalize these insights into a four-layer model—Capacity, Alignment, Levers, and Outcomes—operationalized through a Capacity × Alignment matrix that delineates success and failure regimes in conversational AI. The framework further identifies key design levers (such as anthropomorphism, authority cues, modality, and onboarding strategies) that modulate user perceptions and outcomes.

Grounded in cross-domain empirical evidence, this work makes three primary contributions: (i) it consolidates disparate evaluative constructs into a coherent structure; (ii) it positions accountability and inclusion as first-order dimensions of dialogue quality; and (iii) it provides an actionable framework for the design and evaluation of conversational systems across modalities. In doing so, we shift the focus from generic notions of interaction quality toward a dialogically grounded, user-centric understanding of what makes human–AI conversations effective, trustworthy, and inclusive.

2. Methods

We conducted an integrative, structured, cross-domain literature review focused on user-driven dimensions of interaction quality in human–AI conversations [22–24]. Our objectives were to (i) identify and systematise user-centric evaluative dimensions reported

in empirical studies of conversational agents and (ii) synthesise these findings to develop an interpretive framework of human–AI dialogue quality. The review encompasses both text-based and voice-based agents, including LLM-powered systems, -like assistants, as well as multimodal or embodied conversational systems, when explicitly addressing interaction quality.

Studies were included if they met the eligibility criteria summarised in Table 1. Applying these criteria to the 2017–2025 window, we retained 125 primary empirical studies. Earlier seminal and review works were tracked for background purposes only and were excluded from the analysed corpus.

Table 1. Eligibility and exclusion criteria used in study selection.

Criterion	Operational Definition
Focus on conversational interaction	Conversational agents/chatbots/dialogue systems (text or voice; including LLM-based); human users are active participants.
User-centric quality dimension	At least one user-driven dimension explicitly described, measured, or discussed (e.g., usability, trust, satisfaction, engagement, empathy, usefulness, fairness, transparency/interpretability, accessibility/inclusion).
Empirical evaluation with users	Lab/field studies, usability tests, surveys/questionnaires, interviews, or focus groups addressing interaction quality.
Publication type	Peer-reviewed journals; full conference/workshop papers; peer-reviewed book chapters/reports.
Language	Full text available in English.
Timeframe	Primary focus on 2017–2025; earlier seminal studies noted as outside the primary window.
Borderline & reviews	Borderline studies retained as secondary evidence (flagged); reviews used only for background or snowballing.
Exclusions	Grey literature (e.g., theses, non-peer-reviewed white papers, non–peer-reviewed sources); machine-to-machine scenarios.

Abbreviation: LLM, large language model.

Rather than employing a fully systematic review protocol characterised by exhaustive query logging and count-based screening (e.g., Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-style), this review adopted a concept-driven, integrative approach. The objective was not to compile an exhaustive inventory of all studies on conversational agents, but to develop a theoretically informed and analytically coherent synthesis of user-centric interaction quality dimensions across domains and agent types.

Accordingly, the literature search was conducted iteratively across multiple databases and publisher platforms, combining structured queries, targeted venue exploration, and backward and forward snowballing. At each iteration, candidate studies were evaluated against the inclusion and exclusion criteria reported in Table 1, and retained if they provided empirical, user-centred evidence relevant to interaction quality in human–AI conversations.

Due to the iterative and integrative nature of this strategy, the review did not rely on a fixed initial retrieval set followed by sequential numerical exclusion steps. Intermediate counts of retrieved and excluded records were therefore not systematically documented, as the process prioritised conceptual relevance and theoretical saturation over exhaustive coverage or record-level traceability.

In line with these criteria, we excluded grey literature, including theses, white papers, and non–peer-reviewed sources, to ensure peer-reviewed quality, and omitted machine-to-machine scenarios, as the focus was strictly on human-centred interaction. No application

domain was excluded, such as healthcare, education, customer service, or public services, provided that user-centric interaction quality was evaluated.

To ensure a structured and multi-source sampling of the relevant literature, we queried major bibliographic databases and digital libraries spanning multiple disciplines. Specifically, we searched *Scopus*, *Web of Science Core Collection*, and *PubMed/MEDLINE* for broad cross-domain coverage, complemented by discipline-specific repositories such as the *ACM Digital Library* and *IEEE Xplore* for HCI and computer science venues. We also conducted targeted searches on publisher platforms (*SpringerLink*, *ScienceDirect/Elsevier*, *Taylor & Francis Online*, *Wiley Online Library*, *JMIR Publications*, and *Frontiers*) to capture recent studies in health informatics, education, and applied AI contexts. All retrieved records were cross-verified via *CrossRef/digital object identifier (DOI)* to confirm peer-reviewed status and bibliographic completeness. This multi-source strategy was adopted to mitigate the coverage bias inherent in any single database and to ensure the inclusion of both journal articles and conference proceedings relevant to the evaluation of user-driven dimensions in human–AI dialogue, without aiming for exhaustive coverage in the manner of a systematic review.

Search terms combined agent-related keywords including *chatbot*, *conversational agent*, *dialogue system*, *voice assistant*, *LLM* with user-centric evaluative constructs including *usability*, *trust*, *satisfaction*, *engagement*, *usefulness*, *explainability/transparency*, *social presence/anthropomorphism*, *accessibility/inclusion*, *fairness*. Reference list checks and snowballing from relevant reviews were used to identify additional eligible studies; such reviews served only as background or sourcing aids, not as primary evidence.

Screening was conducted in two stages—initial title and abstract screening, followed by full-text eligibility assessment based on the criteria outlined above—within the iterative, integrative search process described earlier. Duplicates were removed before screening. Borderline cases were retained as secondary sources and flagged to avoid over-weighting. Reviews were logged only for snowballing and theoretical context.

Across the screening stages, exclusions primarily resulted from a limited and recurrent set of reasons aligned with the eligibility criteria in Table 1. At the title and abstract level, studies were most commonly excluded for not involving conversational agents or human-facing interaction (e.g., agent–agent or system-internal dialogue), not addressing user-centric interaction quality, or being non-empirical in nature (e.g., conceptual discussions or technical system descriptions without user evaluation). At the full-text stage, exclusions typically reflected ineligible publication types (e.g., theses, white papers, or non–peer-reviewed sources), the absence of explicit user-centred quality dimensions, or insufficient methodological detail to support reliable data extraction. Studies falling outside the temporal or language constraints were also excluded at this stage. This qualitative accounting clarifies how the inclusion and exclusion criteria were applied in practice within an integrative, iterative review process, rather than through a count-based systematic screening pipeline.

In analysing the eligible studies, we systematically noted key descriptive and evaluative elements that guided the synthesis, summarised in Table 2. These elements included agent type and modality, domain and context, study design and participant profile, user-centric dimensions and measures employed, user-perceived outcomes, and salient design factors. They served as analytical foci for structuring the comparative narrative rather than as a basis for exhaustively tabulating all studies.

Table 2. Data items extracted from eligible studies (examples are illustrative, not exhaustive).

Item	Definition/Examples
Agent type & modality	Text, voice, multimodal; LLM-augmented where applicable.
Domain/context	Application area (e.g., healthcare, education, customer service, public services).
Study design & participants	Controlled lab, field, usability test, survey; sample size and profile.
User-centric dimensions	Operational definitions and constructs assessed.
Measures/instruments	Instruments/scales (e.g., SUS, UEQ, CUQ, BUS-15/CUS) and bespoke items.
User-perceived outcomes	Usefulness/task effectiveness, trust/credibility, engagement/hedonic quality, satisfaction, acceptance/continuance.
Design factors	Repair strategies, timing/synchronicity, persona/authority cues, explanation/transparency, accessibility/inclusion affordances.

Abbreviations: SUS, System Usability Scale; UEQ, User Experience Questionnaire; CUQ, Chatbot Usability Questionnaire; BUS-15, BOT Usability Scale (15 items); CUS, Chatbot Usability Scale.

Extraction and coding were conducted using an iteratively refined codebook. An initial coding pass identified candidate dimensions, constructs, and grouping rules based on the extracted items in Table 2. Subsequent passes refined category definitions, harmonised labels across studies, and consolidated semantically overlapping constructs (e.g., usefulness vs. perceived utility; transparency vs. explainability), while preserving distinctions specific to conversational interaction (e.g., breakdown handling and repair, turn-taking timing, onboarding, and disclosure).

At the outset of the analysis, an initial discussion with co-authors served to align the conceptual scope and high-level category structure. Thereafter, extraction and coding were carried out by a single reviewer. Ambiguous cases and overlapping constructs were treated as intra-coder ambiguities and resolved through iterative comparison with operational definitions, re-examination of the original study context, and consistency checks across the corpus, with category boundaries refined accordingly. Priority was given to interpretations that were most consistently supported across studies and aligned with explicit user-reported measures. Coding decisions and borderline cases were documented throughout the process to support internal consistency and minimise interpretive drift.

As the review aimed at interpretive synthesis rather than effect estimation, the coding scheme was used to structure the comparative narrative and inform framework construction, rather than to produce exhaustive frequency counts.

We employed a narrative, comparative synthesis to identify cross-cutting regularities and tensions across domains. Three families of regularities emerged: a pragmatic core, including usability, task effectiveness, and conversational competence; a social-affective layer, including social presence, warmth, and perceived synchronicity; and an accountability and inclusion layer, including explanation efficacy, honesty/disclosure, source transparency, and accessibility/fairness. These regularities, together with observed moderators and mediating mechanisms, guided the development of a multi-layer interpretive framework and the Capacity × Alignment matrix, which links levers to outcomes.

We used a VOSviewer (version 1.6.20) co-occurrence map as a triangulation aid; details and interpretation are provided in Section 4.

Building on this synthesis, we organized the user-driven dimensions into four layers—*Capacity*, *Alignment*, *Levers/Moderators*, and *Outcomes*—to offer a structured interpretation of how users assess *human–AI dialogue quality*. In this framework, dialogue quality emerges from the enactment of technical competence (Capacity) and contextual appropriateness (Alignment) through specific levers, and how this interplay shapes user-perceived outcomes. To operationalise this construct analytically, we projected *Capacity* and *Alignment*

onto orthogonal axes, resulting in a *Capacity* × *Alignment matrix* that delineates distinct success and failure regimes (e.g., high-capacity but misaligned interactions; well-aligned yet under-capable systems). The matrix functions as an interpretive device for reasoning about how specific levers—such as anthropomorphism and social presence, authority cues and role framing, modality and voice characteristics, onboarding and guidance, and tone or personalisation—can shift user perceptions toward high-quality dialogue within a given contextual setting.

Because constructs, measures, and contexts are heterogeneous (including varied user experience (UX) scales and domain-specific instruments), we did not conduct a meta-analysis. Instead, we focused on mechanistic and relational consistencies, such as mediation through social presence and trust calibration via explanation/disclosure. Finally, although the review spans multiple domains and modalities, publication bias and uneven reporting, including the absence of reliability statistics or construct definitions in some studies, remain potential threats to the completeness and generalizability of the findings. Our framework should therefore be understood as an empirically grounded, context-sensitive model rather than a prescriptive industry taxonomy.

3. Findings

This section presents a thematic synthesis of the analysed literature, organised around recurrent user-centred dimensions of interaction quality. We selectively report representative empirical evidence to illustrate each dimension.

3.1. Usability, Ease of Use, and Conciseness

Usability—understood as the ease and intuitiveness with which users can interact with a conversational agent—emerges as a foundational dimension of interaction quality across domains [1,25]. From the user's perspective, usability encompasses not only task completion but the overall interaction journey, including clarity of language, navigational effort, and perceived smoothness of dialogue, all of which contribute to engagement and sustained adoption [10,11,26].

Within this pragmatic core, conciseness has been repeatedly identified as a key determinant of perceived ease of use. Evidence from voice-based and accessibility-focused studies shows that short, information-dense responses are often judged to be as useful as longer utterances, while being more efficient and cognitively manageable—particularly for command-oriented tasks and eyes-free interaction [9,27]. At the same time, conciseness is not a universal proxy for quality: experimental work with LLM-powered agents indicates that reducing conversational length does not consistently improve perceived interaction quality across tasks, suggesting that brevity should be adaptive to user goals and context rather than uniformly minimised [5]. Taken together, these findings position conciseness as a key usability attribute—one that supports efficiency and reduces cognitive load while allowing for on-demand depth when required.

Empirical evaluations across various application domains confirm that usability is a prerequisite for effective conversational systems. High ease-of-use ratings and strong usability scores have been reported for health-oriented chatbots supporting Coronavirus disease 2019 (COVID-19) screening, diabetes self-management, and rehabilitation, with users frequently attributing positive experiences to clear guidance and intuitive interaction flows [4,28–30]. For example, in health-screening chatbots, users often report positive usability experiences in terms of completing interactions without uncertainty about the next steps, noting that "there is no way you can get lost" due to clear instructions and prompts [4]. More broadly, high usability scores have also been reported for diabetes

self-management chatbots, with SUS results indicating strong perceived ease of use and navigability [28,29].

However, reliance on generic instruments such as the System Usability Scale (SUS) has also revealed limitations. Several studies show that while SUS and the User Experience Questionnaire (UEQ) often yield high scores, they may fail to capture interaction-specific aspects of dialogue-based systems, motivating the development of chatbot-specific measures such as the Chatbot Usability Questionnaire (CUQ) and the Chatbot Usability Scale (CUS) within the BOT Usability Scale (15 items) (BUS-15) [16,31,32].

Overall, the literature converges on a view of usability as a multifaceted construct that extends beyond interface clarity to include responsiveness, conversational efficiency, and perceived effort during interaction. These attributes are consistently linked to acceptance and continued use, as ease of use shapes behavioural intention in technology acceptance models and, conversely, usability breakdowns often lead to frustration and abandonment [1,7,33]. Importantly, several studies caution that broad labels such as "ease of use" may obscure meaningful distinctions from the user's perspective, underscoring the need for more granular, interaction-specific descriptors when evaluating conversational AI [16].

3.2. Usefulness, Task Effectiveness, and Conversational Competence

Beyond ease of use, users consistently evaluate conversational agents in terms of usefulness—that is, their ability to effectively support goal achievement. Perceived usefulness, a core construct in technology acceptance models, alongside task success, emerges as a central determinant of user satisfaction and overall interaction quality across domains [16,34,35]. From the user's perspective, interaction quality is strongly shaped by whether the agent provides efficient assistance and successfully resolves the problem, with failures in understanding or task completion being a primary source of frustration [10].

Empirical studies operationalise this dimension through task completion rates, perceived success, and self-reported usefulness. Classic frameworks such as PARADISE conceptualise user satisfaction as a function of task success and interaction cost, while more recent evaluations commonly rely on users' judgments of how effectively the agent addressed their needs [16,36]. Across application domains, higher perceived usefulness and task effectiveness consistently correlate with increased satisfaction, trust, and acceptance [35,37].

A recurring finding in this literature is that perceived usefulness is not determined solely by task outcomes, but is critically mediated by conversational competence. Studies consistently show that agents capable of managing breakdowns—through acknowledgment, repair strategies, and the provision of actionable alternatives—reduce interactional costs and improve perceived task success [38,39]. Specifically, when a chatbot fails to understand a request, users report markedly different evaluations depending on how the breakdown is handled. Agents that explicitly acknowledge the misunderstanding and offer actionable paths forward, such as reformulation options, are often still perceived as helpful. In contrast, strategies that proceed without signalling the breakdown—delivering likely incorrect responses—or that rely on simple repetition-based repairs are consistently associated with lower perceived usefulness and more negative user evaluations. This illustrates how conversational repair practices directly influence users' judgments of task effectiveness.

Effective turn-taking and response timing also influence perceived effectiveness: adaptive timing benefits different user groups, whereas misaligned conversational pacing often necessitates repair and can jeopardise task completion [8,40]. During failures, communicative strategies such as apologies and warm verbal or vocal styles have been shown to mitigate negative reactions and sustain users' willingness to continue the interaction, thereby preserving perceived usefulness despite errors [41,42].

Evidence from domain-specific studies reinforces the primacy of competence over superficial interactional features. In both healthcare and customer service contexts, design choices related to language, guidance, persona, and interface structure significantly influence perceived effectiveness and user outcomes, whereas anthropomorphic cues alone do not compensate for limited task performance [43–45]. Large-scale surveys in e-commerce similarly show that perceived understanding and problem-resolution ability are decisive in users' willingness to rely on chatbots, while personality-related attributes matter only insofar as they are grounded in demonstrable competence [45].

Taken together, these findings underscore that usefulness and task effectiveness form the pragmatic core of interaction quality: when conversational agents fail to support users' goals accurately and efficiently, other positive aspects of the interaction become secondary.

3.3. Creativity, Diversity, and Originality

Beyond task effectiveness, conversational agents can influence how broadly and inventively users explore the solution space. An emerging body of empirical work therefore examines creativity-related dimensions of interaction quality from a user perspective, including perceived generativity, originality, novelty, and idea diversity.

Across ideation and brainstorming contexts, studies consistently show that interacting with conversational agents can enhance perceived generativity and the breadth of ideas produced. Experimental findings indicate that users collaborating with chatbot partners tend to generate a greater number of ideas and report higher perceived idea quality compared to human-only or control conditions, suggesting that the presence of an AI collaborator can lower inhibitions and stimulate divergent thinking [46,47]. More recent work on LLM-augmented ideation further shows that conversational agents can expand the outcome space and support creative flow when integrated into structured processes—both individual and group brainstorming—provided that interaction pacing and turn-taking are carefully managed [48,49]. At the individual level, comparative studies also suggest that chatbots can perform competitively in divergent thinking tasks, often exceeding average human originality scores while remaining complementary rather than dominant creative partners [50].

At the same time, creativity-related benefits are not unconditionally positive at the collective level. Evidence from group settings indicates that while conversational agents may increase idea quantity, reliance on a shared AI partner can reduce overall idea diversity—highlighting risks of homogenisation when many users draw from the same generative source [6].

Taken together, these findings underscore a critical evaluation challenge: creativity-enhancing effects may vary across individual and collective contexts. Assessments of interaction quality should therefore account not only for perceived novelty and generativity at the individual level, but also for potential trade-offs in diversity and variance at the group level.

3.4. User Satisfaction, Overall Experience, and Inclusivity

User satisfaction is commonly treated as an overarching outcome of interaction quality, reflecting users' global appraisal of their experience with a conversational agent. Across studies, satisfaction is typically measured through post-interaction self-report and is strongly associated with users' perceptions of whether the system fulfilled its intended purpose and provided a smooth, acceptable interaction [4,16]. As such, satisfaction functions as a synthetic indicator that integrates multiple facets of interaction quality, rather than as an isolated construct.

Empirical evidence consistently shows that satisfaction is shaped by both functional and relational determinants. Successful problem resolution, accuracy, and efficiency emerge as primary drivers, while responsiveness, conversational quality, and perceived warmth further influence users' overall evaluations [10,51,52]. Notably, satisfaction does not always align directly with task success: users may report positive experiences despite incomplete outcomes when agents adopt supportive or apologetic communication styles, whereas frictional or frustrating interaction processes can lead to dissatisfaction even when goals are technically achieved [53,54]. These findings indicate that satisfaction reflects a joint appraisal of both outcomes and the interactional process.

The determinants of satisfaction also vary across application domains. In healthcare, education, and customer service contexts, satisfaction may encompass domain-specific expectations such as feeling supported, learning effectively, or experiencing reduced waiting times. It often mediates longer-term outcomes including trust, loyalty, and continued use [55–57]. This domain sensitivity underscores the role of satisfaction as both an immediate experiential judgment and a bridge toward sustained human–AI relationships.

Importantly, satisfaction is closely intertwined with inclusivity, accessibility, and fairness. When conversational agents exhibit systematic performance disparities or fail to accommodate diverse linguistic, cognitive, or physical needs, users' overall evaluations tend to decline, and interaction breakdowns may be perceived as exclusionary or even discriminatory [58]. Conversely, inclusive design choices—such as multilingual support, multimodal interaction, and transparent feedback—have been shown to improve perceived usability, conversational competence, and satisfaction, particularly among vulnerable or underserved user groups [59–61].

Taken together, these findings position inclusivity not as a peripheral concern, but as a central determinant of user satisfaction and overall interaction quality.

3.5. Trust, Perceived Credibility, and User Agency

As conversational agents increasingly take on roles traditionally associated with human service providers, trust has emerged as a central dimension of interaction quality from the user's perspective [12]. Trust is commonly defined as a user's willingness to accept vulnerability to an agent's actions based on expectations of competent, reliable, and benevolent performance. In practice, higher trust is associated with greater reliance on the agent's advice, increased information disclosure, and a higher likelihood of selecting the agent over human alternatives, whereas low trust often leads to avoidance, verification behaviours, or abandonment [12,35].

Across empirical studies, trust is shaped by a constellation of functional, interactional, and contextual factors. Core antecedents include perceived competence and reliability; transparency regarding system capabilities and limitations; and assurances of security and privacy. Social and conversational cues—such as politeness, warmth, and responsiveness—also contribute to users' trust assessments [12,53]. Experimental evidence consistently shows that explanation efficacy and clarity enhance perceived credibility and trust, while modality-related cues—such as voice versus text or vocal characteristics—can further influence users' credibility judgments and their willingness to rely on the agent [62–64].

Importantly, trust in conversational agents is dynamic rather than static. While it can be undermined by errors or failures, it can also be repaired through effective communicative strategies, such as acknowledging mistakes, offering apologies, and adopting socially supportive interaction styles [41,53]. At the same time, evidence from high-stakes domains suggests that excessive displays of simulated empathy may backfire, and that communicative competence and epistemic appropriateness are often more critical for sustaining trust than overt human-likeness [65]. For instance, in travel- or service-oriented

chatbots, empirical studies show that users report higher trust when the agent briefly explains the rationale behind a recommendation (e.g., cost, timing, or user preferences), compared to interactions where suggestions are presented without justification [62]. In such cases, trust arises not from perceived intelligence alone, but from the agent's ability to make its reasoning transparent within the dialogue—consistent with broader evidence on explanation-driven trust in conversational AI [66].

Beyond agent behaviour alone, recent studies highlight the role of user agency and control in shaping trust. Providing users with mechanisms for oversight, control, and expectation management—such as uncertainty disclosures, limitation statements, or opportunities to intervene—has been shown to increase appropriate trust and reduce overreliance without imposing additional cognitive burden [67,68]. Disclosure of an agent's identity or degree of human involvement can recalibrate trust in context-dependent ways, sometimes reducing immediate credibility but increasing tolerance and understanding following failures [69,70].

Together, these findings position trust not merely as a passive attitude but as an interactional outcome co-shaped by system behaviour, communicative design, and the degree of agency afforded to users.

3.6. Engagement, Enjoyment, and Hedonic Quality

Engagement refers to the degree of user involvement, immersion, and willingness to continue interacting with a conversational agent, while enjoyment (or hedonic quality) captures the positive affective experience derived from the interaction. Together, these dimensions underscore that interaction quality is not determined solely by functional performance—particularly in consumer-facing or non-task-oriented contexts—but also by the extent to which the interaction is perceived as pleasant, stimulating, or socially rewarding [71,72].

Empirical work consistently distinguishes between pragmatic quality (e.g., usefulness, efficiency) and hedonic quality (e.g., enjoyment, novelty, stimulation), showing that both contribute to overall user experience, albeit through different mechanisms. Studies in service and retail contexts indicate that affective design features—such as humour, informal language, personality cues, and conversational warmth—can enhance enjoyment and perceived social presence, which in turn increase satisfaction and willingness to engage with conversational agents [3,45,73]. Importantly, enjoyment often functions as a mediator: social or human-like cues may not directly increase usage intentions but do so indirectly by making the interaction more enjoyable.

At the same time, evidence points to trade-offs and boundary conditions. Design choices that enhance hedonic quality—such as topic-led dialogue or playful interaction styles—may reduce perceived efficiency or pragmatic performance. Their effectiveness depends on contextual factors, including task criticality, failure severity, and brand or domain expectations [3,73]. These findings suggest that engagement-enhancing strategies must be carefully calibrated rather than universally applied.

In evaluation studies, engagement is commonly operationalised through users' intentions to continue, return to, or recommend use, as well as through self-reported enjoyment and behavioural indicators such as sustained interaction. Evidence further suggests that the importance of hedonic quality varies across user groups and contexts: younger users and leisure-oriented scenarios tend to place greater emphasis on enjoyment and playfulness, whereas utilitarian or high-stakes settings prioritise task completion and efficiency [2,10].

Taken together, these findings position engagement and enjoyment as context-sensitive yet essential components of interaction quality, shaping both immediate user experience and longer-term acceptance and usage.

3.7. Anthropomorphism and Social Presence

Anthropomorphism refers to users' perceptions of a conversational agent as human-like in appearance or behaviour, while social presence captures the subjective feeling of interacting with a social other rather than a machine. Both are inherently user-centric dimensions, reflecting how users experience the agent's persona and the social quality of the interaction. Prior work consistently shows that these perceptions can shape interaction quality by influencing engagement, trust, and enjoyment, although poorly aligned human-like cues may also lead to unmet expectations or discomfort [74,75].

Across empirical studies, anthropomorphism primarily exerts its effects through social presence. Human-like linguistic and interactional cues—such as conversational warmth, personality, or socially oriented dialogue—tend to increase perceived social presence, which in turn enhances downstream outcomes including trust, satisfaction, and engagement across service, retail, and public-sector contexts [3,76,77]. This mediating role suggests that anthropomorphism is rarely beneficial in isolation; its impact depends on whether it successfully fosters a sense of social connection during interaction.

At the same time, evidence highlights clear boundary conditions. Anthropomorphic cues can backfire when they raise expectations the agent cannot meet, leading to expectancy violations, frustration, or reduced credibility. Studies consistently show that superficial human-like features (e.g., names or visual identity cues) are insufficient on their own and may even be detrimental if not supported by adequate conversational competence and responsiveness [78,79]. Conversely, richer interactional cues can compensate for minimal visual anthropomorphism, underscoring the importance of aligning perceived social qualities with actual system capabilities.

From an evaluation perspective, anthropomorphism and social presence contribute positively to interaction quality when they are moderate, context-appropriate, and aligned with the agent's functional abilities. Well-calibrated human-like cues can facilitate more natural and engaging interactions by enhancing social presence, whereas excessive or misaligned anthropomorphism risks undermining trust and acceptance. Taken together, these findings position anthropomorphism and social presence as conditional enablers of interaction quality rather than universally desirable design goals.

3.8. Acceptance and Continued Use Intentions

A central expectation of high interaction quality is that a positive user experience translates into acceptance, understood as users' willingness to adopt a conversational agent for regular use and to integrate it into their routines or tasks. Accordingly, many empirical studies operationalise acceptance through behavioural intentions, such as the intention to use or reuse the agent, willingness to recommend it, or preference for the chatbot over alternative interaction modalities. Acceptance is critical from a user-centred perspective, as even technically capable systems offer limited value if users are unwilling to engage with them over time.

Across domains, evidence consistently shows that acceptance is shaped by a small set of core perceptions. Models derived from the Technology Acceptance Model (TAM), the Unified Theory of Acceptance and Use of Technology (UTAUT), and related frameworks converge in identifying perceived usefulness, ease of use, and trust as primary antecedents of behavioural intention—often mediated by overall attitude or satisfaction. Empirical studies in insurance, education, healthcare, and cross-cultural settings show that trust frequently exerts the strongest influence, followed by usefulness and usability. This suggests that improving specific interaction quality dimensions can incrementally increase acceptance, even when baseline willingness is low [35,80,81].

Acceptance also manifests as a comparative preference between interaction alternatives. Studies in service and e-commerce contexts show that users are more willing to switch from human agents to chatbots when the latter are perceived as competent, reliable, and enjoyable, whereas service failures or low perceived understanding often lead users to revert to human support. Affective and social cues—including anthropomorphic or warm communication styles—can mitigate rejection following failures, but do not compensate for persistent deficiencies in task performance [45,82].

Overall, these findings reinforce that acceptance is a composite outcome, grounded in both functional and experiential dimensions of interaction quality.

Beyond initial adoption, continued use and retention depend on expectation-confirmation processes. Research drawing on such models consistently shows that satisfaction and perceived usefulness drive continuance intention, with both influenced by whether prior expectations are met or exceeded during interaction. This mechanism has been observed across healthcare, service, and mental health applications, and extends to recent studies of LLM-based assistants, where high perceived value supports adoption, while concerns about correctness, misuse, or domain suitability temper acceptance in high-stakes contexts [7,83–85].

Taken together, acceptance and continued use intentions emerge as downstream, integrative indicators of interaction quality rather than independent design targets.

3.9. Explanation and Transparency Efficacy

An increasingly salient user-centric dimension of interaction quality concerns the perceived efficacy of explanations provided by conversational agents. From a user perspective, explanation efficacy reflects whether explanations support understanding, confidence, and appropriate reliance during interaction—rather than merely exposing system logic. Across empirical studies, effective explanations are consistently associated with higher perceived transparency, trust, and acceptance [62,66].

Selective experimental evidence shows that explanations improve interaction quality when they are intelligible, context-sensitive, and aligned with users' needs. Studies comparing explanation-rich versus explanation-poor conversational agents consistently report improvements in users' understanding, trust, and willingness to rely on recommendations when explanations are provided—particularly when users lack prior domain knowledge. Dialogic and socially framed explanations further enhance perceived transparency and satisfaction by enabling users to actively engage in sense-making, rather than passively receiving justifications [86,87].

Importantly, transparency extends beyond explanatory content to include honest and responsible disclosure about the agent's identity, limitations, and information sources. Users interpret transparency broadly, encompassing clarity about data practices, the provenance of information, and the socio-technical context of interaction. Empirical work shows that identity disclosure, source transparency, and explanations of system behaviour—such as response delays or uncertainty—can recalibrate expectations and foster trust, particularly in situations involving errors or ambiguity. However, such disclosures may reduce perceived competence in high-stakes contexts if not carefully framed [69,70,88,89].

Overall, these findings indicate that explanation and transparency efficacy is not determined solely by the completeness of technical content, but by its communicative and epistemic alignment with users' mental models and goals. From an interaction-quality perspective, transparency functions as a situated, user-centred communicative act that supports understanding, trust calibration, and user agency. Poorly designed or misleading explanations, by contrast, risk confusion, false confidence, or erosion of trust—underscoring

the need to evaluate explanation efficacy in terms of user understanding and decision quality, rather than the mere presence of explanatory features [90,91].

These themes are integrated in the next section through a comparative synthesis that identifies cross-cutting patterns, moderators, and mediating mechanisms, and provides the foundation for the interpretive framework.

4. Comparative Synthesis

This section consolidates the evidence presented in Section 3 into a comparative synthesis. Rather than proposing a formal model, it surfaces cross-cutting empirical regularities—recurring patterns, tensions, moderators, and mediating mechanisms—that span domains and modalities. These regularities provide the empirical foundation for the interpretive framework developed in the following section.

4.1. Cross-Cutting Patterns and Emergent Clusters

At the corpus level, the pragmatic core consistently co-occurs with design choices that manage cognitive load and conversational costs. Adaptive concision—short outputs expandable on demand—supports eyes-free use and reduces effort, but length itself is not a monotonic proxy for quality in LLM-mediated dialogue. Instead, brevity must be calibrated to task goals, channel constraints, and user profiles [5,9,27]. Task success remains the dominant benchmark for usefulness and satisfaction, while the micro-dynamics of repair and turn-taking materially shape perceived effectiveness: acknowledging misunderstandings, proposing actionable alternatives, escalating to humans when necessary, and timing responses to user expertise reliably improve outcomes [10,37–40,92,93].

Layered on this core, social–affective cues enhance evaluations when they are aligned with the context. Warmth, apologetic tone, and selective use of humour can buffer service failures and increase social presence, yielding downstream gains in satisfaction and re-engagement; these benefits depend on factors such as failure severity, brand equity, and user expectations [53,73,94]. Perceived competence and credibility are further shaped by modality and authority signals: vocal and accent cues, role framing including doctor vs. student, and perceived expertise influence trust and willingness to defer, particularly in high-stakes domains [35,43,63,64,95,96]. In parallel, explanation and transparency practices—including layered rationales, identity and limitation disclosures, and verifiable provenance—improve understanding, calibrate reliance, and support acceptance. At the same time, the literature warns against 'placebic' explanations and credibility inflation from unverified citations [62,66,87,91,97–99]. Finally, inclusivity and accessibility function as first-order quality determinants: accounting for dialectal variation and code-mixing, supporting motor and cognitive differences, and enabling efficient eyes-free navigation directly enhance perceived ease, satisfaction, and willingness to continue [9,58–60,100].

A transversal theme concerns creativity. At the individual level, conversational AI often expands the idea space—boosting perceived generativity and novelty—yet at the collective level, it may compress diversity when many participants draw from the same model or prompting pattern. Recent group studies highlight the risk of homogenization unless variance-preserving mechanisms including multi-persona prompting, explicit divergence phases, are integrated into the interaction [6,46,48–50].

In parallel, hedonic and motivational outcomes—such as enjoyment, flow, and continuance intention—are jointly shaped by pragmatic value and perceived social presence [7,35,45,72,80,81,84].

To triangulate these themes, we computed a VOSviewer co-occurrence map using titles and abstracts, with a threshold of three. The resulting communities align with the clusters discussed above: usability and user experience, social and affective factors and

anthropomorphism, transparency and repair, adoption and acceptance, and modality and context. We present this as corpus-level corroboration rather than as a taxonomy. In the map, node size indicates term frequency, edge thickness indicates co-occurrence strength, and colours indicate modularity-based communities (Figure 1).

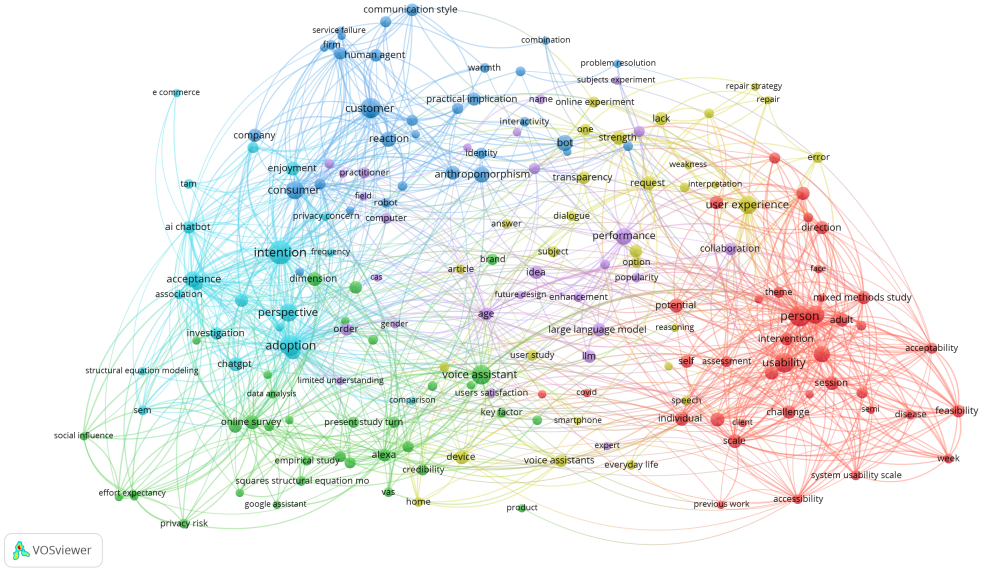

Figure 1. VOSviewer co-occurrence map of the review corpus (titles and abstracts; threshold = 3).

4.2. Tensions, Trade-Offs, and Moderators

The findings reveal several empirically grounded tensions. Concision improves efficiency and accessibility but can under-serve complex goals if not expandable. Topic-led, human-like dialogue increases social presence and enjoyment but may reduce pragmatic efficiency. Anthropomorphic cues can enhance engagement yet also raise expectations such that insufficient capability or interactivity triggers expectancy violations [3,5,27,78,79]. Transparency poses a similar calibration challenge: identity and limitation disclosures strengthen integrity and tolerance for errors but, if poorly framed, can reduce perceived competence and willingness to reuse in high-stakes contexts [69]. Personalisation introduces its own tension with privacy and security, which can be mitigated through provenance, verifiability, and meaningful user controls [98,99]. Finally, while synchronicity and speed are valued in service contexts, short delays—when explained—can improve perceived transparency and trust, indicating that temporal behaviour itself is a communicative design lever [45,89].

These dynamics are moderated by domain risk, modality, and user profile. High-stakes domains such as health and finance raise expectations for trust, disclosure, and provenance relative to retail or entertainment [35,95,96]. Voice-specific features, including accent, shift credibility and persuasion; persona and institutional role affect usability and willingness to comply; and user attributes such as expertise, age, region, and native language shape timing preferences and barrier experiences [2,10,40,43,63,64]. Human-in-the-loop configurations and their disclosure recalibrate expectations and reliance in situ, offering actionable levers for trust calibration [70,95]. In group ideation, the balance shifts between productivity and diversity, underscoring how context systematically shapes quality outcomes [6,48].

4.3. Mediating Mechanisms and Measurement Implications

Across models and domains, a small set of mediation paths appears consistently. Anthropomorphic cues typically act through social presence to influence trust, sat-

isfaction, and behavioural intentions. Explanation and transparency practices improve understanding, which in turn supports calibrated trust and acceptance. Usefulness and satisfaction reliably mediate continuance intention in technology-adoption paradigms [62,66,76,77,83,84,87,101]. Warmth and tone similarly operate through social presence to facilitate post-failure recovery [53,73].

These mechanisms have direct implications for measurement. Generic UX instruments such as SUS, UEQ, capture baseline usability but miss dialogue-specific nuances. Chatbot-tailored instruments such as CUQ, BUS-15/CUS, more directly assess conversational efficiency, information quality, privacy/security, and response time [16,31]. Evaluation batteries should align multi-level outcomes: objective or perceived task success; usefulness, trust, and satisfaction; inclusion and accessibility indicators (e.g., dialectal error disparities); and behavioural traces such as retention or voluntary return use. For explanation and transparency, subjective ratings of clarity and understanding should be paired with verifiability checks to guard against placebo effects and unearned credibility. In creativity contexts, measures of idea quantity and novelty should be complemented with diversity at both individual and group levels [6,91]. Not least, longitudinal designs remain scarce despite their importance for modeling trust dynamics and continuance over time [12].

5. Interpretive Framework

The comparative synthesis reveals recurring regularities that motivate the need for a structured framework capable of integrating diverse dimensions into a coherent account of evaluation. To systematise these findings, we propose a four-layer interpretive framework that organises dimensions into *Capacity*, *Alignment*, *Levers/Moderators*, and *Outcomes*. Together, these layers articulate the emergent construct of *human–AI dialogue quality*. In other words, human–AI dialogue quality is not treated as a fixed system attribute, but as a relational construct that manifests through dialogue, as *Capacity* and *Alignment* are enacted, moderated, and ultimately reflected in user-centred outcomes.

At the foundation lies *Capacity*, defined as the agent's technical ability to generate appropriate outputs and fulfil user needs in line with its designed purpose. Dimensions mapped to this layer include usability and conciseness, conversational competence (repair, turn-taking, and handover), accuracy, relevance, and fluency, as well as domain knowledge, responsiveness and latency, and creativity, originality, and generativity. These capture the system's raw functional potential—the extent to which it can act as a competent conversational partner.

The second layer, *Alignment*, denotes the degree to which the system's behaviour is consistent with user goals, expectations, constraints, and ethical standards. It encompasses explanation and transparency (including source transparency), honesty and ethical responsibility, safety and harm avoidance, inclusivity, accessibility, and fairness, as well as user agency and trust calibration including disclosure, uncertainty, oversight, and privacy and data security. Whereas capacity describes what the system can do, alignment reflects whether and how it does so appropriately in context.

The third layer, *Levers and Moderators*, identifies the design mechanisms that shape how capacity and alignment are enacted and perceived in dialogue. These include anthropomorphism and social presence, authority cues and role framing, modality and voice/accent, guidance and onboarding strategies, and personalisation of persona and tone including humour, empathy. As the synthesis showed, such levers can amplify trust, engagement, and satisfaction when consistent with capacity and context, but may also create expectancy violations or credibility inflation when misaligned.

The final layer, *Outcomes*, captures the user's perceived results of the interaction. These include usefulness and task effectiveness, satisfaction and overall experience, trust and

perceived credibility, engagement and hedonic quality, and acceptance and continuance. Outcomes represent the observable metrics by which dialogue quality is ultimately judged.

To make explicit the dynamic interplay between the first two layers, we map capacity and alignment onto orthogonal axes, yielding the *Capacity* × *Alignment matrix* (Figure 2). This surface illustrates four distinct zones:

- *Sweet spot (high capacity, high alignment)*: systems deliver adoption, trust, and positive outcomes.
- *Over-power, under-guard (high capacity, low alignment)*: technically accurate but unsafe, unfair, or opaque systems risk over-trust and harmful reliance.
- *Well-behaved but unhelpful (low capacity, high alignment)*: agents are safe and transparent but ineffective, leading to frustration and abandonment.
- *Rejection (low capacity, low alignment)*: systems offer little value and are not adopted.

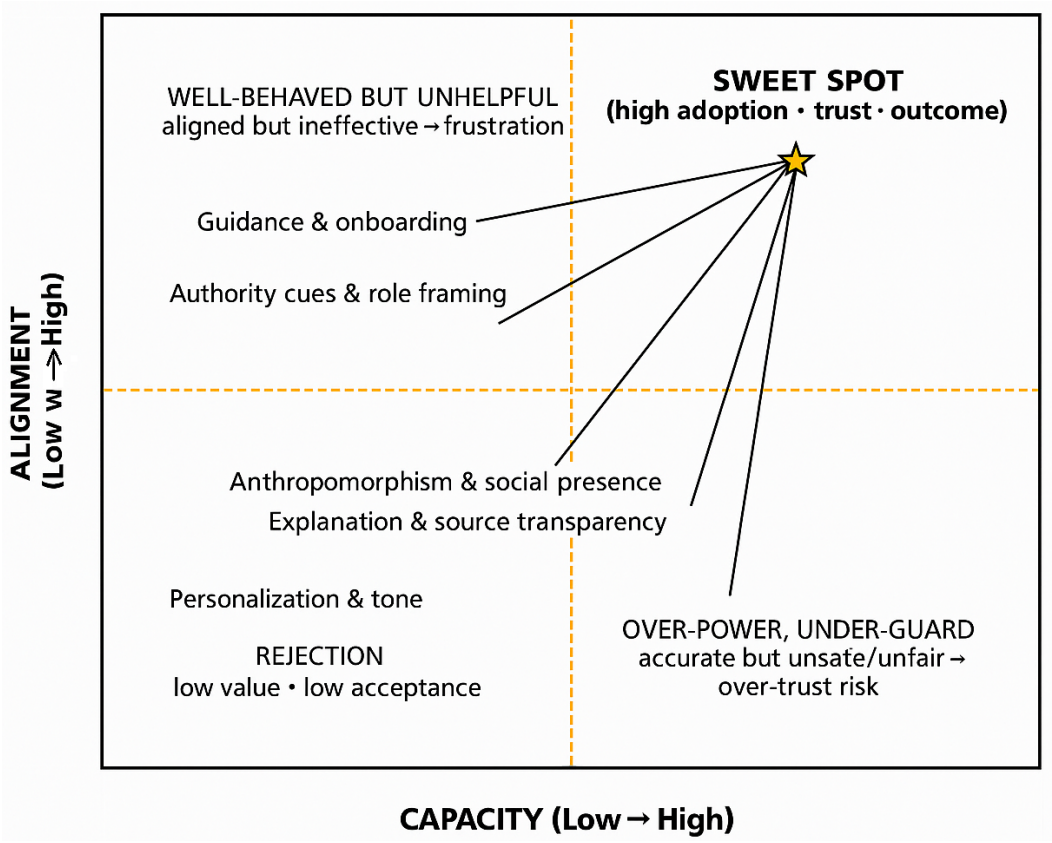

Figure 2. Capacity × Alignment matrix with lever-driven nudge paths. Arrows indicate the direction of lever-driven nudges (i.e., how design levers can shift perceived dialogue quality across regimes), while the yellow star marks the target 'sweet spot' (high capacity, high alignment). The surface illustrates success and failure regimes and clarifies the consequences of unresolved gaps (e.g., over-trust when capacity outpaces alignment; abandonment when alignment is adequate but capacity insufficient).

To illustrate these regimes more concretely, consider a health symptom-checker that delivers fluent and confident recommendations but fails to disclose uncertainty or limitations. In such cases, users may over-rely on its outputs despite a misalignment with clinical risk (high *Capacity*, low *Alignment*). Conversely, an overly cautious agent that consistently foregrounds uncertainty and safety constraints, but repeatedly refuses to offer actionable guidance—even when safe alternatives are available—may be perceived as trustworthy yet ineffective, leading to user frustration and eventual abandonment (low *Capacity*, high *Alignment*).

Levers such as explanation, anthropomorphism, authority cues, and personalisation function as 'nudge paths' within this matrix, shifting user perceptions toward or away from the sweet spot. The matrix operationalises the key tensions surfaced in the synthesis: verbosity versus conciseness, warmth versus efficiency, disclosure versus competence, and personalisation versus privacy.

Levers should be understood as regulatory mechanisms that mediate how capacity and alignment translate into outcomes. Anthropomorphism and social presence can foster trust and engagement, but only when interactivity sustains user expectations. Authority cues and role framing may increase compliance, yet risk inappropriate deference if not accompanied by transparency and oversight. Explanation and source transparency can calibrate reliance but may also inflate credibility when unverified. These dynamics underscore the need to design levers not as superficial add-ons but as integral components of dialogue governance.

Taken together, the synthesis suggests that existing notions of interaction quality are insufficient to account for the dialogic mechanisms through which users evaluate conversational agents in practice.

Building on this evidence, we introduce the construct of *human–AI dialogue quality (HADQ)*, defined as the emergent property of interactions in which Capacity and Alignment are co-orchestrated through specific levers, producing measurable user outcomes:

$$\text{HADQ} = f(\text{Capacity}, \text{Alignment}; \text{Levers}) \longrightarrow \text{Outcomes}.$$

This expression is not meant as a mathematical model to be estimated, but rather as a conceptual shorthand that makes explicit the relational logic of the framework: dialogue quality arises from the joint enactment of Capacity and Alignment in context, mediated by specific levers, and becomes observable through user-centred outcomes.

This conceptual move is warranted for three reasons. First, it provides conceptual precision by shifting the focus from the generic notion of interaction quality to the specific dynamics of dialogue. Second, it ensures empirical adequacy, as findings consistently show that users evaluate quality through dialogic mechanisms such as turn-taking, repair, explanation, and tone. Third, it offers diagnostic utility, since the Capacity $\times$ Alignment matrix and lever-driven pathways are only meaningful when dialogue itself is recognised as the primary locus of evaluation.

By foregrounding the co-orchestration of Capacity and Alignment through levers, and by positioning outcomes as user-anchored evaluative criteria, the interpretive framework advances a precise, empirically grounded, and actionable understanding of what constitutes quality in human–AI dialogue. This reconceptualisation recomposes the fragmented landscape into a coherent construct that maps current debates, surfaces tensions and gaps, and clarifies the stakes of leaving these gaps unresolved.

6. Discussion and Implications

Building on the dialogical synthesis developed in this review, this section outlines the implications of the findings for theory, design, evaluation, and governance.

From a theoretical perspective, the results suggest that established adoption and user experience models capture only the downstream manifestations of conversational interaction. Constructs such as perceived usefulness, satisfaction, and continuance emerge not as primary drivers, but as outcomes of how conversational capacity and contextual alignment are enacted during dialogue. In other words, users do not evaluate isolated system properties in the abstract, but form their judgments based on the quality of the dialogue as it unfolds over time. By foregrounding dialogue as the locus of evaluation, the HADQ construct complements existing models by clarifying where and how trust, engagement, and reliance are formed—and where they break down.

For design, the synthesis highlights pragmatic conversational competence as a necessary baseline for perceived quality. Low-effort navigation, relevance, and effective handling of breakdowns are prerequisites upon which social–affective cues and transparency mechanisms can build. When a system's self-presentation exceeds its actual capabilities or contextual fit, systematic failure regimes emerge—including frustration, abandonment, and inappropriate reliance. These patterns indicate that high-leverage design mechanisms—such as adaptive concision, explicit role framing, calibrated expressions of uncertainty [102,103], and repair-oriented dialogue strategies—should be treated as first-order design decisions rather than peripheral refinements. For example, in student support settings, explicitly framing the agent as a "teaching assistant" that cites relevant lecture sources and provides brief, expandable answers can reduce interactional cost (Capacity) while calibrating user reliance through provenance cues and explicit uncertainty management (Alignment). By contrast, a highly personable agent that implies authoritative expertise without verifiable references may inflate perceived credibility and prompt inappropriate deference when errors occur. Inclusivity and accessibility also emerge as integral to dialogue quality, as robustness to linguistic variation, cognitive differences, and modality constraints directly shapes perceived usability and trust.

Implications for evaluation follow directly from these observations. While generic UX instruments offer a useful baseline, they remain insufficient for capturing dialogic phenomena such as repair, explanation efficacy, trust calibration, and social presence. Chatbot-specific measures, behavioural traces (e.g., clarification turns, refusals, escalation events), and longitudinal indicators of continued use are necessary to assess how dialogue quality evolves over time. Recent standardisation efforts in adjacent fields—such as the Artificial Social Agent (ASA) questionnaire developed within the Intelligent Virtual Agent community—demonstrate the feasibility of coordinated measurement practices beyond ad hoc study instruments [104]. However, consistent with the patterns identified in this review, subjective ratings alone are insufficient to assess calibrated trust and should be complemented with behavioural and verifiability-oriented indicators.

Finally, the findings underscore the need to treat dialogue quality as a governance-relevant, risk-sensitive construct. Clear specification of intended use, target users, and known failure regimes is essential to prevent inappropriate reliance. Transparency-by-design, traceable logging, and institutionalised human oversight should be calibrated to domain-specific risk, with stronger safeguards in high-stakes contexts. Longer-term risks-including automation-induced deskilling and the erosion of human judgment [105]-further motivate explicit consent mechanisms, lifecycle management, and periodic re-certification, including clearly defined constraints on personalisation and memory.

In summary, high-quality human–AI dialogue does not emerge from isolated improvements in usability, explainability, or governance. It depends on the coherent orchestration of conversational capacity, contextual alignment, and risk-aware oversight. By making these dynamics explicit, the HADQ construct provides both a focused theoretical lens and a practical vocabulary for designing, evaluating, and governing conversational agents as reliable and trustworthy partners across domains.

7. Research Agenda

Building on the integrative synthesis and the interpretive framework developed in this review, several avenues for future research become evident. These directions are necessary to consolidate the construct of human–AI dialogue quality (HADQ), extend its empirical grounding, and refine its methodological utility across domains.

- Existing evidence relies primarily on short-term studies with convenience samples, providing only a limited view of how capacity, alignment, and user outcomes evolve

over time. Future work should conduct longitudinal and cross-contextual field studies—particularly in high-stakes domains—to track how trust, satisfaction, and continuance intentions develop across repeated interactions and under both successful and failed conditions.

- Current measurement tools either capture only generic usability or remain undervalidated across domains and cultures. There is a need for standardised, multidimensional batteries that integrate generic and dialogue-specific metrics, incorporate accessibility and inclusion indicators, and triangulate subjective reports with behavioural traces such as dropout rates, voluntary return use, and—where feasible—physiological markers of engagement. While recent community-driven efforts have demonstrated the feasibility and value of coordinated, standardised evaluation instruments for artificial social agents, they also underscore the challenges of applying fixed questionnaires across diverse agent paradigms. In particular, LLM-based conversational systems present distinct issues related to dialogic grounding, epistemic alignment, and accountability—issues that call for integrative frameworks to guide the principled selection, adaptation, and interpretation of evaluation tools across contexts.
- The effects of design levers such as anthropomorphism, authority cues, and personalisation are highly context-dependent and insufficiently explored. Future research should examine how domain risk, task criticality, user characteristics, and cultural factors moderate these effects, using controlled experiments and cross-cultural comparisons to determine when these levers enhance HADQ and when they undermine it.
- Conversational AI can enhance individual creativity but may reduce diversity at the group level due to homogenised model outputs or convergent prompting patterns. Future studies should design and test variance-preserving mechanisms—such as multi-persona prompting, divergent ideation scaffolds, and structured turn-taking—to balance productivity with diversity in collective settings.
- Transparency, disclosure, fairness, and related governance levers remain theoretically emphasised yet empirically under-tested, and long-term risks such as deskilling are poorly understood. Future empirical trials should embed governance mechanisms (e.g., uncertainty displays, provenance cues, escalation logs) and evaluate their effects on calibrated reliance, error tolerance, perceived responsibility, and the evolution of human judgement over time.
- Insights relevant to HADQ remain fragmented across disciplines and are only loosely connected to emerging regulatory frameworks and technical standards. Interdisciplinary collaborations should refine core constructs, strengthen empirical grounding, and align evaluation methods with frameworks such as the European Union Artificial Intelligence Act (EU AI Act) and International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) standards. In parallel, computational models linking interaction traces, user modelling, and outcome prediction could support proactive safeguards within the Capacity $\times$ Alignment framework.

In sum, advancing this agenda requires moving beyond fragmented evaluations toward a cumulative science of human–AI dialogue quality. By situating capacity and alignment as co-determinants of user experience, and by treating levers as regulatory mechanisms rather than superficial add-ons, the framework introduced here provides a foundation for systematic inquiry. Pursuing the outlined directions will refine theoretical constructs and empirical measures and consolidate HADQ as a unifying paradigm for studying, comparing, and governing conversational agents across research domains and real-world practice.

8. Limitations

This review is subject to several limitations that qualify the scope and interpretation of its contributions. First, the synthesis is intentionally user-driven: we prioritised dimensions of dialogue quality reported in empirical studies centred on users' experiences and perceptions. While this focus aligns with our goal of modelling human–AI dialogue quality, it de-emphasises system-internal metrics including model perplexity, latency under load, and organisational outcomes such as workflow efficiency or safety indicators, unless explicitly tied to user-facing constructs. The framework may therefore underrepresent factors that influence quality indirectly through technical or organisational layers.

Second, the evidence base is highly heterogeneous. Studies differ in domain, modality, population, and measurement instruments, often employing non-standard scales or idiosyncratic operationalisations. Because of this variability, we did not conduct a meta-analysis; instead, we emphasised mechanistic and relational consistencies. While this supports theory-building, it limits claims about the magnitude or relative importance of dimensions. Moreover, many included studies rely on short-term, self-reported outcomes from laboratory or pilot deployments, which are vulnerable to recall and novelty biases and provide limited ecological validity. Few combine perceptual measures with behavioural logs or longitudinal follow-up, leaving longer-term dynamics underexplored. In addition, the breadth of our cross-domain synthesis—spanning text and voice modalities and low- to high-stakes contexts—may come at the expense of actionable specificity. This highlights the need for domain- and modality-specific operationalisations to complement the general framework.

Third, the construction of mapping introduces interpretive subjectivity. Despite iterative coding and consensus procedures, boundaries between conceptually adjacent dimensions (e.g., usefulness vs. task effectiveness; clarity vs. concision; trust vs. calibrated reliance) can blur. Our Capacity $\times$ Alignment matrix discretises what is in practice a continuous space: while this aids clarity, it may obscure gradients and interactions present in real dialogue. Similarly, designating levers as moderators reflects interpretive choices grounded in the reviewed evidence rather than causal identification. More broadly, the four-layer model and the Capacity $\times$ Alignment matrix have not yet been prospectively validated through experimental or field studies; future work should assess their predictive validity and decision-usefulness in situated deployments.

Fourth, coverage across contexts and populations is uneven. The literature is skewed toward peer-reviewed studies and well-resourced domains such as HCI, education, consumer services, with limited evidence from underrepresented languages, accessibility scenarios including screen-reader users, cognitive or sensory differences, and high-stakes or safety-critical applications. Generalizability to cross-cultural settings and vulnerable populations should therefore be treated with caution. Moreover, while our approach was structured and multi-source, it does not constitute a systematic review in the PRISMA sense. Consequently, the absence of record-level screening counts and a PRISMA-style flow diagram reflects a deliberate methodological choice aligned with an integrative, concept-driven synthesis, rather than an omission or lack of transparency. The emphasis was on conceptual integration rather than exhaustive coverage, which may have left some relevant studies uncaptured despite mitigation strategies.

In line with this, some recent and highly relevant domain-specific studies may not appear in the analysed corpus, despite being fully consistent with the stated eligibility criteria. For example, recent work on co-designed or stakeholder-inclusive embodied conversational agents in healthcare and rehabilitation contexts (e.g., [106,107]) illustrates interaction-quality dimensions—such as relational experience, engagement, and inclusivity—that align closely with the framework proposed here. These studies were not captured

during the iterative search and screening process and were therefore not included in the coded corpus. Their absence reflects the integrative and non-exhaustive nature of the review, rather than any intentional exclusion, and they serve to corroborate—rather than challenge—the generality and applicability of the synthesised dimensions. Such examples are cited solely to illustrate the external resonance and plausibility of specific interaction-quality dimensions and were not used as primary evidence in the extraction, coding, or synthesis process.

In addition, our inclusion criteria—limited to English-language, peer-reviewed sources-and the exclusion of grey literature may introduce selection biases, potentially underrepresenting user-centric evidence emerging from local, practitioner-oriented, or non-English venues. We also note that unreviewed preprints were excluded from the analysed corpus to ensure the inclusion of peer-reviewed studies; as a result, some very recent or emerging evidence may not be represented until it is formally published.

Fifth, the technological landscape is evolving rapidly. The baseline Capacity of foundation models and conversational stacks changes quickly; advances in reasoning, retrieval, and tool use can shift failure regimes and the salience of specific dimensions within short timeframes. Findings anchored in earlier model vintages or non-instrumented agents may not transfer directly to newer systems with richer capabilities such as structured tool use, uncertainty displays, multi-agent orchestration. Relatedly, our treatment of LLM-specific challenges for instance hallucination/fabrication, prompt sensitivity and instruction-following variability, alignment via reinforcement learning from human feedback (RLHF) and its failure modes, and agentic behaviours with external tools, is necessarily brief in a cross-domain review and warrants dedicated measurement protocols and reporting standards.

Finally, additional threats to validity remain. The proposed framework is explanatory rather than causal: while the pathways synthesise recurrent patterns, experimental evidence is needed to establish causal mechanisms. Publication bias, variability in study quality, and incomplete reporting of sampling or instrument validation may also shape which dimensions appear most consistently.

We mitigated these risks by emphasising convergent regularities across heterogeneous studies, clearly indicating areas where evidence remains sparse or inconsistent, and refraining from causal claims where data did not warrant them. Thus, the limitations delineate the scope of our synthesis without undermining its substantive contributions. They signal priority areas for further research while positioning the framework as a robust foundation for refinement, validation, and domain-specific operationalisation.

9. Conclusions

This review set out to address the fragmented state of research on interaction quality in human–AI conversations. Prior work often treated quality as a narrow set of usability or performance metrics, leaving a conceptual gap in how user experiences, ethical considerations, and dialogic mechanisms jointly shape evaluations. To address this gap, we synthesised evidence across domains to conceptualise human–AI dialogue quality (HADQ) as an emergent construct grounded in the co-orchestration of system capacity and contextual alignment via design levers.

Our findings show that users consistently evaluate dialogue quality across three interrelated layers: a pragmatic core comprising usability, task effectiveness, and conversational competence; a social–affective layer encompassing social presence, warmth, and engagement; and an accountability and inclusion layer emphasising transparency, fairness, and accessibility. These insights informed a four-layer interpretive framework—Capacity, Alignment, Levers, and Outcomes—operationalised through the Capacity × Alignment matrix. This model not only delineates success and failure regimes but also provides diagnostic

value for guiding design and governance decisions. Notably, the review underscores that levers such as anthropomorphism, tone, and personalisation function not merely as aesthetic features but as regulatory mechanisms with measurable impacts on trust, satisfaction, and user acceptance.

The implications of this work are both practical and theoretical. For designers, the framework provides a roadmap for striking a balance between pragmatic efficiency and ethical responsibility, enabling the development of conversational agents that are usable, trustworthy, and inclusive. For evaluators and regulators, it provides a vocabulary and set of tools for benchmarking systems against multidimensional standards rather than isolated metrics. More broadly, HADQ reframes dialogue quality as a relational and dialogical phenomenon—users assess not only what agents deliver, but also how they do so, and whether this process aligns with values such as competence, fairness, and accountability.

At the same time, several limitations temper these contributions. The evidence base remains heterogeneous in terms of domains, populations, and measurement instruments, with many studies relying on short-term, self-reported outcomes. While our synthesis emphasises regularities across contexts, it cannot establish causal mechanisms or precise effect sizes. Moreover, the rapid evolution of conversational AI suggests that findings based on earlier systems may not fully capture the dynamics of emerging foundation models. These constraints, however, do not diminish the framework's value as a conceptual foundation; rather, they identify priorities for future validation and contextual extension.

Future research should focus on longitudinal and cross-contextual studies to investigate how trust, satisfaction, and reliance evolve over time. Advances in measurement science are needed to triangulate subjective judgments with behavioural data and physiological markers. Empirical trials should further investigate the double-edged effects of design levers, while governance-oriented research must test mechanisms such as uncertainty displays, provenance tracking, and human-in-the-loop oversight. Interdisciplinary collaborations will be essential for aligning HADQ with regulatory frameworks and for developing predictive models that anticipate failure regimes before they escalate.

In closing, this work calls for a decisive shift: from fragmented assessments of interaction quality to a cumulative science of human–AI dialogue quality. By recognising capacity and alignment as co-determinants of user experience—and by treating design levers as regulatory mechanisms rather than superficial features—the framework offers a durable foundation for developing conversational agents that are not only competent but also responsible and human-centred.

Author Contributions: Conceptualization, L.M. and F.C.; methodology, L.M. and F.C.; formal analysis, L.M.; investigation (literature search, screening and coding), L.M.; data curation, L.M.; visualization, L.M. and L.L.; writing—original draft preparation, L.M., L.L. and F.C.; writing—review and editing, L.M. and L.L.; supervision, L.L. and F.C.; project administration, L.M. and F.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Acknowledgments: ChatGPT (OpenAI, version 5.2) was used during manuscript preparation to support language editing and improve clarity and readability of the text. The authors take full responsibility for the content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Milne-Ives, M.; Cock, C.; Lim, E.; Shehadeh, M.; Pennington, N.; Mole, G.; Mole, G.; Meinert, E. The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. *J. Med. Internet Res.* **2020**, *22*, e20346. [CrossRef] [PubMed]
2. Han, S.; Lee, M.K. FAQ chatbot and inclusive learning in massive open online courses. *Comput. Educ.* **2022**, *179*, 104395. [CrossRef]
3. Haugeland, I.K.F.; Følstad, A.; Taylor, C.; Bjørkli, C.A. Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design. *Int. J. Hum.-Comput. Stud.* **2022**, *161*, 102788. [CrossRef]
4. Chagas, B.A.; Pagano, A.S.; Prates, R.O.; Praes, E.C.; Ferreguetti, K.; Vaz, H.; Reis, Z.S.N.; Ribeiro, L.B.; Ribeiro, A.L.P.; Pedroso, T.M.; et al. Evaluating User Experience With a Chatbot Designed as a Public Health Response to the COVID-19 Pandemic in Brazil: Mixed Methods Study. *JMIR Hum. Factors* **2023**, *10*, e43135. [CrossRef]
5. Huang, S.H.; Lin, Y.F.; He, Z.; Huang, C.Y.; Huang, T.H.K. How Does Conversation Length Impact User's Satisfaction? A Case Study of Length-Controlled Conversations with LLM-Powered Chatbots. In *CHI EA '24: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024*; Association for Computing Machinery: New York, NY, USA, 2024. [CrossRef]
6. Lee, B.C.; Chung, J.J. An empirical investigation of the impact of ChatGPT on creativity. *Nat. Hum. Behav.* **2024**, *8*, 1906–1914. [CrossRef]
7. Ma, J.; Wang, P.; Li, B.; Wang, T.; Pang, X.S.; Wang, D. Exploring User Adoption of ChatGPT: A Technology Acceptance Model Perspective. *Int. J. Hum.–Comput. Interact.* **2025**, *41*, 1431–1445. [CrossRef]
8. Porcheron, M.; Fischer, J.E.; Reeves, S.; Sharples, S. Voice Interfaces in Everyday Life. In *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–12. [CrossRef]
9. Vtyurina, A.; Fourney, A.; Morris, M.R.; Findlater, L.; White, R.W. VERSE: Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search. In *ASSETS '19: Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, Pittsburgh, PA, USA, 28–30 October 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 414–426. [CrossRef]
10. Følstad, A.; Brandtzaeg, P.B. Users' experiences with chatbots: Findings from a questionnaire study. *Qual. User Exp.* **2020**, *5*, 3. [CrossRef]
11. Følstad, A.; Taylor, C. Investigating the user experience of customer service chatbot interaction: A framework for qualitative analysis of chatbot dialogues. *Qual. User Exp.* **2021**, *6*, 6. [CrossRef]
12. Ng, S.W.T.; Zhang, R. Trust in AI chatbots: A systematic review. *Telemat. Inform.* **2025**, *97*, 102240. [CrossRef]
13. McGuire, J.; Cremer, D.D.; Hesselbarth, Y.; Schutter, L.D.; Mai, K.M.; Hiel, A.V. The reputational and ethical consequences of deceptive chatbot use. *Sci. Rep.* **2023**, *13*, 16246. [CrossRef]
14. Deriu, J.; Rodrigo, A.; Otegi, A.; Echegoyen, G.; Rosset, S.; Agirre, E.; Kübler, S.; Hall, M.; Rodrigo, A.; Müller, T.; et al. Survey on evaluation methods for dialogue systems. *Artif. Intell. Rev.* **2021**, *54*, 755–810. [CrossRef]
15. Jannach, D. Evaluating conversational recommender systems. *Artif. Intell. Rev.* **2023**, *56*, 2365–2400. [CrossRef]
16. Borsci, S.; Malizia, A.; Schmettow, M.; Bernhaupt, T.; Blackwell, C.; Ma, A. The Chatbot Usability Scale: The Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Pers. Ubiquitous Comput.* **2022**, *26*, 95–119. [CrossRef]
17. Pupo, L.G.H.; Herrera, R.Y.; Acuña, L.A.; Pérez, P.Y.P.; Pupo, I.P.; Pérez, R.B. Measurement of Perceived Quality in Conversational Systems (Chatbots). In *Computational Intelligence Applied to Decision-Making in Uncertain Environments*; Pérez, P.Y.P., Pupo, I.P., Kacprzyk, J., Pérez, R.E.B., Eds.; Studies in Computational Intelligence; Springer: Cham, Switzerland, 2025; Volume 1195, pp. 219–236. [CrossRef]
18. Loveys, K.; Sebaratnam, G.; Sagar, M.; Broadbent, E. The Effect of Design Features on Relationship Quality with Embodied Conversational Agents: A Systematic Review. *Int. J. Soc. Robot.* **2020**, *12*, 1293–1312. [CrossRef]
19. Kuhail, M.A.; Taj, I.; Alimamy, S.; Shawar, B.A. A review on polyadic chatbots: Trends, challenges, and future research directions. *Knowl. Inf. Syst.* **2025**, *67*, 109–165. [CrossRef]
20. de Filippis, M.L.; Federici, S.; Mele, M.L.; Borsci, S.; Bracalenti, M.; Gaudino, G.; Cocco, A.; Amendola, M.; Simonetti, E. Preliminary Results of a Systematic Review: Quality Assessment of Conversational Agents (Chatbots) for People with Disabilities or Special Needs. In *Computers Helping People with Special Needs. ICCHP 2020*; Miesenberger, K., Manduchi, R., Rodriguez, M.C., Peňáz, P., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12376, pp. 343–350. [CrossRef]
21. Namvarpour, M.M.; Razi, A. The Art of Talking Machines: A Comprehensive Literature Review of Conversational User Interfaces. In *CUI '25: Proceedings of the 7th ACM Conference on Conversational User Interfaces, Waterloo, ON, Canada, 8–10 July 2025*; Association for Computing Machinery: New York, NY, USA, 2025. [CrossRef]
22. Snyder, H. Literature review as a research methodology: An overview and guidelines. *J. Bus. Res.* **2019**, *104*, 333–339. [CrossRef]
23. Grant, M.J.; Booth, A. A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Inf. Libr. J.* **2009**, *26*, 91–108. [CrossRef]

24. Torraco, R.J. Writing Integrative Literature Reviews: Guidelines and Examples. *Hum. Resour. Dev. Rev.* **2005**, *4*, 356–367. [CrossRef]
25. Bennion, M.; Hardy, G.; Moore, R.; Kellett, S.; Millings, A. Usability, Acceptability, and Effectiveness of Web-Based Conversational Agents to Facilitate Problem Solving in Older Adults: Controlled Study. *J. Med. Internet Res.* **2020**, *22*, e16794. [CrossRef]
26. Jain, M.; Kumar, P.; Kota, R.; Patel, S.N. Evaluating and Informing the Design of Chatbots. In *DIS '18: Proceedings of the 2018 Designing Interactive Systems Conference, Hong Kong, China, 17–19 August 2018*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 895–906. [CrossRef]
27. Haas, G.; Rietzler, M.; Jones, M.; Rukzio, E. Keep it Short: A Comparison of Voice Assistants' Response Behavior. In *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022*; Association for Computing Machinery: New York, NY, USA, 2022. [CrossRef]
28. Swallow, V.; Horsman, J.; Mazlan, E.; Campbell, F.; Zaidi, R.; Julian, M.; Branchflower, J.; Martin-Kerry, J.; Monks, H.; Soni, A.; et al. DigiBete, a Novel Chatbot to Support Transition to Adult Care of Young People/Young Adults With Type 1 Diabetes Mellitus: Outcomes From a Prospective, Multimethod, Nonrandomized Feasibility and Acceptability Study. *JMIR Diabetes* **2025**, *10*, e74032. [CrossRef]
29. Bruijnes, M.; Kesteloo, M.; Brinkman, W.P. Reducing social diabetes distress with a conversational agent support system: A three-week technology feasibility evaluation. *Front. Digit. Health* **2023**, *5*, 1149374. [CrossRef]
30. Hocking, J.; Maeder, A.; Powers, D.; Perimal-Lewis, L.; Dodd, B.; Lange, B. Mixed methods, single case design, feasibility trial of a motivational conversational agent for rehabilitation for adults with traumatic brain injury. *Clin. Rehabil.* **2024**, *38*, 322–336. [CrossRef] [PubMed]
31. Holmes, S.; Moorhead, A.; Bond, R.; Zheng, H.; Coates, V.; McTear, M. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *ECCE '19: Proceedings of the 31st European Conference on Cognitive Ergonomics, Belfast, UK, 10–13 September 2019*; ACM, Association for Computing Machinery: Belfast, UK, 2019; pp. 207–214. [CrossRef]
32. de Souza Cavalcanti, A.R.; do Nascimento, J.W.A.; da Silva Lima Roque, G.; de Souza, R.R.; de Melo Queiroz, S.R.; Corrêa, J.A. A Virtual Assistant in Vaccine Pharmacovigilance: Content and Usability Validation. *CIN Comput. Inform. Nurs.* **2023**, *41*, 482–490. [CrossRef]
33. Fergencs, T.; Meier, F. Engagement and Usability of Conversational Search—A Study of a Medical Resource Center Chatbot. In *Diversity, Divergence, Dialogue, 16th International Conference, iConference 2021, Beijing, China, 17–31 March 2021*; Toeppe, K., Yan, H., Chu, S.K.W., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 328–345.
34. Venkatesh, V.; Davis, F.D. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Manag. Sci.* **2000**, *46*, 186–204. [CrossRef]
35. de Andrés-Sánchez, J.; Gené-Albesa, J. Not with the bot! The relevance of trust to explain the acceptance of chatbots by insurance customers. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 110. [CrossRef]
36. Walker, M.; Litman, D.; Kamm, C.A.; Abella, A. PARADISE: A framework for evaluating spoken dialogue agents. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, 7–12 July, 1997; pp. 271–280.
37. Rese, A.; Tränkner, P. Perceived conversational ability of task-based chatbots—Which conversational elements influence the success of text-based dialogues? *Int. J. Inf. Manag.* **2024**, *74*, 102699. [CrossRef]
38. Ashktorab, Z.; Jain, M.; Liao, Q.V.; Weisz, J.D. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Scotland, UK, 4–9 May 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–12. [CrossRef]
39. Braggaar, A.; Verhagen, J.; Martijn, G.; Liebrecht, C. Conversational Repair Strategies to Cope with Errors and Breakdowns in Customer Service Chatbot Conversations. In *Chatbot Research and Design, 7th International Workshop, CONVERSATIONS 2023, Oslo, Norway, 22–23 November 2024*; Følstad, A., Araujo, T., Papadopoulos, S., Law, E.L.C., Luger, E., Goodwin, M., Hobert, S., Brandtzaeg, P.B., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 23–41.
40. Gnewuch, U.; Morana, S.; Adam, M.T.P.; Maedche, A. Opposing Effects of Response Time in Human–Chatbot Interaction. *Bus. Inf. Syst. Eng.* **2022**, *64*, 773–791. [CrossRef]
41. de Sá Siqueira, M.A.; Müller, B.C.N.; Bosse, T. When Do We Accept Mistakes from Chatbots? The Impact of Human-Like Communication on User Experience in Chatbots That Make Mistakes. *Int. J. Hum.–Comput. Interact.* **2024**, *40*, 2862–2872. [CrossRef]
42. Huang, B.; Sénécal, S. How should voice assistants be heard? The mitigating effect of verbal and vocal warmth in voice assistant service failure. *Serv. Ind. J.* **2023**, *43*, 806–826. [CrossRef]
43. Biro, J.; Linder, C.; Neyens, D. The Effects of a Health Care Chatbot's Complexity and Persona on User Trust, Perceived Usability, and Effectiveness: Mixed Methods Study. *JMIR Hum. Factors* **2023**, *10*, e41017. [CrossRef] [PubMed]

44. Nguyen, M.H.; Sedoc, J.; Taylor, C.O. Usability, Engagement, and Report Usefulness of Chatbot-Based Family Health History Data Collection: Mixed Methods Analysis. *J. Med. Internet Res.* **2024**, *26*, e55164. [CrossRef]
45. Chen, S.; Wang, P.; Wood, J. Exploring the varying effects of chatbot service quality dimensions on customer intentions to switch service agents. *Sci. Rep.* **2025**, *15*, 22559. [CrossRef]
46. Wieland, B.; de Wit, J.; de Rooij, A. Electronic Brainstorming With a Chatbot Partner: A Good Idea Due to Increased Productivity and Idea Diversity. *Front. Artif. Intell.* **2022**, *5*, 880673. [CrossRef]
47. Hwang, A.H.C.; Won, A.S. IdeaBot: Investigating Social Facilitation in Human-Machine Team Creativity. In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021*; Association for Computing Machinery: New York, NY, USA, 2021. [CrossRef]
48. Shaer, O.; Cooper, A.; Mokryn, O.; Kun, A.L.; Ben Shoshan, H. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024*; Association for Computing Machinery: New York, NY, USA, 2024. [CrossRef]
49. Muller, M.; Houde, S.; Gonzalez, G.; Brimijoin, K.; Ross, S.I.; Moran, D.A.S.; Weisz, J.D. Group Brainstorming with an AI Agent: Creating and Selecting Ideas. In *Proceedings of the International Conference on Computational Creativity (ICCC 2024), Jönköping, Sweden, 17–21 June 2024*; Association for Computational Creativity: Jyllinge, Denmark, 2024; p. 10.
50. Koivisto, M.; Grassini, S. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Sci. Rep.* **2023**, *13*, 13601. [CrossRef] [PubMed]
51. Kvale, K.; Freddi, E.; Hodnebrog, S.; Sell, O.A.; Følstad, A. Understanding the User Experience of Customer Service Chatbots: What Can We Learn from Customer Satisfaction Surveys? In *Proceedings of the Chatbot Research and Design, Virtual, 23–24 November 2021*; Følstad, A., Araujo, T., Papadopoulos, S., Law, E.L.C., Luger, E., Goodwin, M., Brandtzaeg, P.B., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 205–218.
52. Ruan, Y.; Mezei, J. When do AI chatbots lead to higher customer satisfaction than human frontline employees in online shopping assistance? Considering product attribute type. *J. Retail. Consum. Serv.* **2022**, *68*, 103059. [CrossRef]
53. Cai, N.; Gao, S.; Yan, J. How the communication style of chatbots influences consumers' satisfaction, trust, and engagement in the context of service failure. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 687. [CrossRef]
54. Xu, Y.; Zhang, J.; Deng, G. Enhancing customer satisfaction with chatbots: The influence of communication styles and consumer attachment anxiety. *Front. Psychol.* **2022**, *13*, 2022. [CrossRef]
55. Essel, H.B.; Vlachopoulos, D.; Tachie-Menson, A.; Johnson, E.E.; Baah, P.K. The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. *Int. J. Educ. Technol. High. Educ.* **2022**, *19*, 57. [CrossRef]
56. Hsu, C.L.; Lin, J.C.C. Understanding the user satisfaction and loyalty of customer service chatbots. *J. Retail. Consum. Serv.* **2023**, *71*, 103211. [CrossRef]
57. Marconi, L.; Cabitza, F. Tutor or Mentor. A Comparative Usability Study of AI Roles in Higher Education. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED, Palermo, Italy, 22–26 July 2025*; Cristea, A.I., Walker, E., Lu, Y., Santos, O.C., Isotani, S., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2025; pp. 201–208.
58. Wenzel, K.; Devireddy, N.; Davison, C.; Kaufman, G. Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition. In *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023*; Association for Computing Machinery: New York, NY, USA, 2023. [CrossRef]
59. Bawa, A.; Khadpe, P.; Joshi, P.; Bali, K.; Choudhury, M. Do Multilingual Users Prefer Chat-bots that Code-mix? Let's Nudge and Find Out! *Proc. ACM Hum.-Comput. Interact.* **2020**, *4*, 1–13. [CrossRef]
60. Masina, F.; Orso, V.; Pluchino, P.; Dainese, G.; Volpato, S.; Nelini, C.; Mapelli, D.; Spagnolli, A.; Gamberini, L. Investigating the Accessibility of Voice Assistants With Impaired Users: Mixed Methods Study. *J. Med. Internet Res.* **2020**, *22*, e18431. [CrossRef]
61. Sezgin, E.; Kocaballi, A.B.; Dolce, M.; Skeens, M.; Militello, L.; Huang, Y.; Stevens, J.; Kemper, A.R. Chatbot for Social Need Screening and Resource Sharing With Vulnerable Families: Iterative Design and Evaluation Study. *JMIR Hum. Factors* **2024**, *11*, e57114. [CrossRef]
62. Joshi, R.; Graefe, J.; Kraus, M.; Bengler, K. Exploring the Impact of Explainability on Trust and Acceptance of Conversational Agents—A Wizard of Oz Study. In *Proceedings of the Artificial Intelligence in HCI, Washington, DC, USA, 29 June–4 July 2024*; Degen, H., Ntoa, S., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 199–218.
63. Flavián, C.; Akdim, K.; Casaló, L.V. Effects of voice assistant recommendations on consumer behavior. *Psychol. Mark.* **2023**, *40*, 328–346. [CrossRef]
64. Pycha, A.; Zellou, G. The influence of accent and device usage on perceived credibility during interactions with voice-AI assistants. *Front. Comput. Sci.* **2024**, *6*, 2024. [CrossRef]
65. Seitz, L.; Bekmeier-Feuerhahn, S.; Gohil, K. Can we trust a chatbot like a physician? A qualitative study on understanding the emergence of trust toward diagnostic chatbots. *Int. J. Hum.-Comput. Stud.* **2022**, *165*, 102848. [CrossRef]

66. Woodcock, C.; Mittelstadt, B.; Busbridge, D.; Blank, G. The Impact of Explanations on Layperson Trust in Artificial Intelligence–Driven Symptom Checker Apps: Experimental Study. *J. Med. Internet Res.* **2021**, *23*, e29386. [CrossRef] [PubMed]
67. Benke, I.; Gnewuch, U.; Maedche, A. Understanding the impact of control levels over emotion-aware chatbots. *Comput. Hum. Behav.* **2022**, *129*, 107122. [CrossRef]
68. Degachi, C.; Tielman, M.L.; Al Owayyed, M. Trust and Perceived Control in Burnout Support Chatbots. In *CHI EA '23: Proceedings of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023*; Association for Computing Machinery: New York, NY, USA, 2023. [CrossRef]
69. Mozafari, N.; Weiger, W.H.; Hammerschmidt, M. Trust me, I'm a bot–repercussions of chatbot disclosure in different service frontline settings. *J. Serv. Manag.* **2021**, *33*, 221–245. [CrossRef]
70. Gnewuch, U.; Morana, S.; Hinz, O.; Kellner, R.; Maedche, A. More Than a Bot? The Impact of Disclosing Human Involvement on Customer Interactions with Hybrid Service Agents. *Inf. Syst. Res.* **2024**, *35*, 936–955. [CrossRef]
71. Hassenzahl, M. The hedonic/pragmatic model of user experience. *Towards UX Manif.* **2007**, *10*, 2007.
72. Hassenzahl, M. The Thing and I: Understanding the Relationship Between User and Product. In *Funology 2: From Usability to Enjoyment*; Blythe, M., Monk, A., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 301–313. [CrossRef]
73. Shams, G.; Kim, K.K.; Kim, K. Enhancing service recovery satisfaction with chatbots: The role of humor and informal language. *Int. J. Hosp. Manag.* **2024**, *120*, 103782. [CrossRef]
74. Nass, C.; Moon, Y. Machines and Mindlessness: Social Responses to Computers. *J. Soc. Issues* **2000**, *56*, 81–103. [CrossRef]
75. Følstad, A.; Nordheim, C.B.; Bjørkli, C.A. What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In *Proceedings of the Internet Science, St. Petersburg, Russia, 24–26 October 2018*; Bodrunova, S.S., Ed.; Springer International Publishing: Cham, Switzerland, 2018; pp. 194–208.
76. Munnukka, J.; Talvitie-Lamberg, K.; Maity, D. Anthropomorphism and social presence in Human–Virtual service assistant interactions: The role of dialog length and attitudes. *Comput. Hum. Behav.* **2022**, *135*, 107343. [CrossRef]
77. Konya-Baumbach, E.; Biller, M.; von Janda, S. Someone out there? A study on the social presence of anthropomorphized chatbots. *Comput. Hum. Behav.* **2023**, *139*, 107513. [CrossRef]
78. Go, E.; Sundar, S.S. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Comput. Hum. Behav.* **2019**, *97*, 304–316. [CrossRef]
79. Tsai, W.H.; Chuan, C.H. How Effective Are Anthropomorphic Chatbots? A Study of Consumer-Chatbot Communication in Taiwan. *Int. J. Commun.* **2024**, *18*, 24.
80. Budhathoki, T.; Zirar, A.; Njoya, E.T.; Timsina, A. ChatGPT adoption and anxiety: A cross-country analysis utilising the unified theory of acceptance and use of technology (UTAUT). *Stud. High. Educ.* **2024**, *49*, 831–846. [CrossRef]
81. Tao, W.; Yang, J.; Qu, X. Utilization of, Perceptions on, and Intention to Use AI Chatbots Among Medical Students in China: National Cross-Sectional Study. *JMIR Med. Educ.* **2024**, *10*, e57132. [CrossRef]
82. Lu, Z.; Min, Q.; Jiang, L.; Chen, Q. The effect of the anthropomorphic design of chatbots on customer switching intention when the chatbot service fails: An expectation perspective. *Int. J. Inf. Manag.* **2024**, *76*, 102767. [CrossRef]
83. Li, X.; Xie, S.; Ye, Z.; Ma, S.; Yu, G. Investigating Patients' Continuance Intention Toward Conversational Agents in Outpatient Departments: Cross-sectional Field Survey. *J. Med. Internet Res.* **2022**, *24*, e40681. [CrossRef]
84. Ashfaq, M.; Yun, J.; Yu, S.; Loureiro, S.M.C. I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telemat. Inform.* **2020**, *54*, 101473. [CrossRef]
85. Zhu, Y.; Wang, R.; Pu, C. "I am chatbot, your virtual mental health adviser." What drives citizens' satisfaction and continuance intention toward mental health chatbots during the COVID-19 pandemic? An empirical study in China. *Digit. Health* **2022**, *8*, 20552076221090031. [CrossRef]
86. Pecune, F.; Murali, S.; Tsai, V.; Matsuyama, Y.; Cassell, J. A Model of Social Explanations for a Conversational Movie Recommendation System. In *HAI '19: Proceedings of the 7th International Conference on Human-Agent Interaction, Kyoto, Japan, 6–10 October 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 135–143. [CrossRef]
87. Hernandez-Bocanegra, D.C.; Ziegler, J. Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *Proceedings of the 3rd Conference on Conversational User Interfaces, Virtual Event, 27–29 July 2021*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1–11. [CrossRef]
88. Natatsuka, A.; Iijima, R.; Watanabe, T.; Akiyama, M.; Sakai, T.; Mori, T. Understanding the Behavior Transparency of Voice Assistant Applications Using the ChatterBox Framework. In *Proceedings of the 25th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '22), Limassol, Cyprus, 26–28 October 2022*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 143–159. [CrossRef]
89. Zhang, Z.; Tsiakas, K.; Schneegass, C. Explaining the Wait: How Justifying Chatbot Response Delays Impact User Trust. In *CUI '24: Proceedings of the 6th ACM Conference on Conversational User Interfaces, Luxembourg, 8–10 July 2024*; Association for Computing Machinery: New York, NY, USA, 2024. [CrossRef]

90. Abdulrahman, A.; Richards, D.; Bilgin, A.A. Reason Explanation for Encouraging Behaviour Change Intention. In Proceedings of the AAMAS, Virtual, 3–7 May 2021; pp. 68–77.
91. Eiband, M.; Buschek, D.; Kremer, A.; Hussmann, H. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *CHI EA '19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–6. [CrossRef]
92. Yeh, S.F.; Wu, M.H.; Chen, T.Y.; Lin, Y.C.; Chang, X.; Chiang, Y.H.; Chang, Y.J. How to Guide Task-oriented Chatbot Users, and When: A Mixed-methods Study of Combinations of Chatbot Guidance Types and Timings. In *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022*; Association for Computing Machinery: New York, NY, USA, 2022. [CrossRef]
93. Cuadra, A.; Li, S.; Lee, H.; Cho, J.; Ju, W. My Bad! Repairing Intelligent Voice Assistant Errors Improves Interaction. *Proc. ACM Hum.-Comput. Interact.* **2021**, *5*, 1–24. [CrossRef]
94. Xie, Y.; Liang, C.; Zhou, P.; Jiang, L. Exploring the influence mechanism of chatbot-expressed humor on service satisfaction in online customer service. *J. Retail. Consum. Serv.* **2024**, *76*, 103599. [CrossRef]
95. Huang, D.; Markovitch, D.G.; Stough, R.A. Can chatbot customer service match human service agents on customer satisfaction? An investigation in the role of trust. *J. Retail. Consum. Serv.* **2024**, *76*, 103600. [CrossRef]
96. Wang, C.; Li, Y.; Fu, W.; Jin, J. Whether to trust chatbots: Applying the event-related approach to understand consumers' emotional experiences in interactions with chatbots in e-commerce. *J. Retail. Consum. Serv.* **2023**, *73*, 103325. [CrossRef]
97. Metzger, L.; Miller, L.; Baumann, M.; Kraus, J. Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024*; Association for Computing Machinery: New York, NY, USA, 2024. [CrossRef]
98. Weidmüller, L.; Etzrodt, K.; Engesser, S. Trustworthiness of voice-based assistants: Integrating interlocutor and intermediary predictors. *Publizistik* **2022**, *67*, 625–651. [CrossRef]
99. Tejwani, R.; Moreno, F.; Jeong, S.; Won Park, H.; Breazeal, C. Migratable AI: Effect of identity and information migration on users' perception of conversational AI agents. In Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, 31 August–4 September 2020; pp. 877–884. [CrossRef]
100. Bleakley, A.; Rough, D.; Roper, A.; Lindsay, S.; Porcheron, M.; Lee, M.; Nicholson, S.A.; Cowan, B.R.; Clark, L. Exploring Smart Speaker User Experience for People Who Stammer. In *ASSETS '22: Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility, Athens, Greece, 23–26 October 2022*; Association for Computing Machinery: New York, NY, USA, 2022. [CrossRef]
101. Janson, A. How to leverage anthropomorphism for chatbot service interfaces: The interplay of communication style and personification. *Comput. Hum. Behav.* **2023**, *149*, 107954. [CrossRef]
102. He, G.; Aishwarya, N.; Gadiraju, U. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *IUI '25: Proceedings of the 30th International Conference on Intelligent User Interfaces, Cagliari, Italy, 24–27 March 2025*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 907–924. [CrossRef]
103. Marconi, L.; Cabitza, F. Show and tell: A critical review on robustness and uncertainty for a more responsible medical AI. *Int. J. Med. Inform.* **2025**, *202*, 105970. [CrossRef]
104. Fitrianie, S.; Bruijnes, M.; Li, F.; Abdulrahman, A.; Brinkman, W.P. The artificial-social-agent questionnaire: Establishing the long and short questionnaire versions. In *IVA '22: Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents, Faro, Portugal, 6–9 September 2022*; Association for Computing Machinery: New York, NY, USA, 2022. [CrossRef]
105. Natali, C.; Marconi, L.; Duran, L.D.D.; Cabitza, F. AI-induced Deskilling in Medicine: A Mixed-Method Review and Research Agenda for Healthcare and Beyond. *Artif. Intell. Rev.* **2025**, *58*, 356. [CrossRef]
106. Hopman, K.; Richards, D.; Norberg, M.N. An Embodied Conversational Agent to Support Wellbeing After Injury: Insights from a Stakeholder Inclusive Design Approach. In *Proceedings of the Persuasive Technology, Wollongong, Australia, 10–12 April 2024*; Baghaei, N., Ali, R., Win, K., Oyibo, K., Eds.; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 161–175.
107. Richards, D.; Miranda Maciel, P.S.; Janssen, H. The Co-Design of an Embodied Conversational Agent to Help Stroke Survivors Manage Their Recovery. *Robotics* **2023**, *12*, 120. [CrossRef]