

# From Commands to Conversation: The Transformative Role of AI-Powered Language Technologies in Human–Computer Interaction

Barkha Shakeelahmad Kasab <sup>1</sup>, Dr. Thota Siva Ratna Sai <sup>2</sup>

<sup>1</sup> Research scholar, Computer Science and Engineering, P. K. University, Shivpuri-Madhya Pradesh

<sup>2</sup> Professor, Computer Science and Engineering, P. K. University, Shivpuri-Madhya Pradesh

## Abstract

Artificial intelligence (AI)-driven natural language processing (NLP) is altering the basic logic of human–computer interaction (HCI). Instead of requiring users to learn menus, commands, syntax, or application-specific workflows, contemporary systems increasingly allow people to express goals through ordinary language and receive context-aware, generative responses. This paper provides an integrative narrative review of the impact of AI-driven NLP on HCI, with particular attention to conversational user interfaces, large language model (LLM)-based assistants, voice interfaces, and adaptive interaction. Drawing on foundational NLP research and empirical HCI studies, the review identifies five major effects: increased interaction naturalness, lower interface learning barriers, stronger personalization and context sensitivity, new opportunities for accessibility and multimodal interaction, and a shift from direct manipulation toward collaborative human–AI problem solving. At the same time, these gains introduce persistent challenges involving hallucination, opacity, bias, privacy, overreliance, expectation management, and unequal performance across users and languages. Evidence from chatbot usability research and recent LLM user studies shows that technical fluency alone does not guarantee high-quality interaction; accuracy, user control, conversational repair, transparency, and calibrated trust are equally important. The paper proposes a human-centered framework linking NLP capabilities to HCI outcomes while treating trust, expertise, safety, and contextual fit as moderating conditions. It concludes that the future of NLP-based HCI should not aim merely to make computers sound human, but to make language-mediated interaction more understandable, controllable, inclusive, and effective.

**Keywords:** artificial intelligence; natural language processing; human–computer interaction; conversational user interfaces; large language models; chatbots; user experience; human–AI interaction

## 1. Introduction

Human–computer interaction has traditionally been shaped by the capabilities and constraints of graphical user interfaces (GUIs). Users learned where functions were located, how commands were represented, and which sequence of actions was required to produce a desired result. Natural language processing changes this relationship because the interface can increasingly interpret a user’s intention from ordinary language rather than requiring the user to translate the intention into machine-oriented syntax. This shift is especially visible in conversational agents, voice assistants, generative AI systems, intelligent search, recommendation tools, and AI copilots. In such systems, language is no longer only content processed by the application; it becomes a principal mechanism through which users control, question, instruct, correct, and collaborate with the computer.

The technical basis for this change developed gradually. The Transformer architecture introduced a highly scalable attention-based approach to sequence modelling (Vaswani et al., 2017). BERT then demonstrated the value of deeply bidirectional pre-training for language understanding across a wide range of NLP tasks (Devlin et al., 2019), while GPT-3 showed that very large language models could perform many tasks from instructions and examples without task-specific retraining (Brown et al., 2020). Instruction tuning and reinforcement learning from human feedback further improved the capacity of language models to follow user intent (Ouyang et al., 2022). These advances have moved NLP from a supporting technology inside applications to a visible interaction layer capable of generating open-ended language, maintaining conversational context, reformulating information, and supporting iterative task completion.

From an HCI perspective, however, improved linguistic capability does not automatically produce improved user experience. Human–AI systems are probabilistic, can be wrong in plausible ways, and may behave inconsistently across prompts or contexts. Amershi et al. (2019) therefore argued that AI interfaces require design principles that

address uncertainty, correction, explanation, and adaptation. Conversational systems add further concerns because fluent language strongly affects perceptions of intelligence, trust, personality, and social presence. Clark et al. (2019) found that users evaluate conversational agents through both functional and social expectations, while Følstad et al. (2018) showed that trust in customer-service chatbots depends on interpretation quality, advice quality, human-likeness, professional presentation, privacy, security, and service context.

The emergence of LLM-based interfaces intensifies these issues. A conversational model can support broad tasks and natural turn-taking, but its openness makes failures harder to predict. Recent studies report user dissatisfaction when LLMs misinterpret intent or provide inaccurate information (Kim et al., 2024), evidence of automation complacency and growing reliance during repeated AI use (Qian & Wexler, 2024), and substantial differences in expectations depending on task and embodiment (Kim, Lee, & Mutlu, 2024). Consequently, the central HCI question is not simply whether AI-driven NLP makes interaction more natural. It is how natural-language interaction changes usability, user expectations, task performance, accessibility, trust, control, and the distribution of responsibility between humans and intelligent systems.

This paper addresses that question through an integrative narrative review. Its objectives are to (1) explain the technological developments that have enabled language-centered interfaces; (2) synthesize evidence on how AI-driven NLP influences usability, engagement, productivity, personalization, and accessibility; (3) identify major human-centered risks and limitations; and (4) propose a practical conceptual framework for designing and evaluating NLP-based HCI. Rather than treating conversational AI as a replacement for established interface paradigms, the paper argues for a complementary approach in which natural language, visual structure, system feedback, and user control are deliberately combined.

### Evolution of AI-Driven Natural Language Interaction

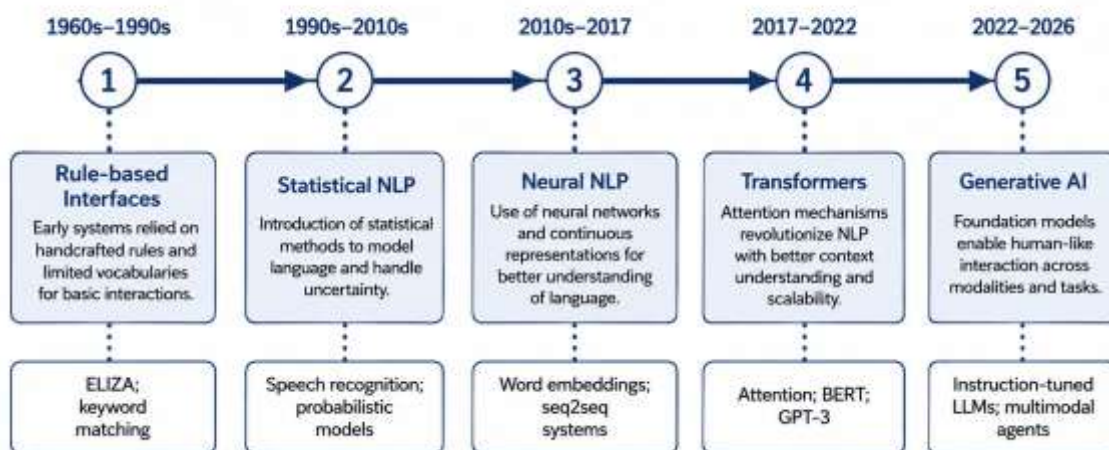


**Figure 1. Evolution from rule-based language interfaces to generative conversational AI.**

## 2. Evolution of AI-Driven NLP in Human-Computer Interaction

The relationship between language technology and HCI predates modern machine learning. Early systems such as ELIZA demonstrated that relatively simple pattern matching could produce the appearance of conversation, but the system had little semantic understanding and depended heavily on scripted responses. Later statistical NLP and speech-recognition systems improved classification, transcription, search, and intent detection, enabling task-oriented dialogue systems with predefined domains and slots. These systems were useful but brittle: users frequently had to discover the phrasing accepted by the interface, and unexpected input often led to failure or fallback messages.

Neural sequence models improved representation learning and language generation, but the Transformer marked a major turning point because self-attention enabled better modelling of long-range dependencies and highly parallel training (Vaswani et al., 2017). BERT strengthened contextual language understanding, while generative transformer models increasingly supported coherent text production across multiple tasks. GPT-3 demonstrated a further interaction shift: rather than selecting from predefined intents, users could specify a task in natural language, provide examples, revise instructions, and obtain generated output within the same conversational interface (Brown et al., 2020). This created a new form of end-user programming in which prompts function as flexible instructions.

The transition from command-based to conversational interaction is therefore not merely a change in input modality. It alters the division of labour between user and system. In a conventional GUI, designers anticipate tasks and encode available operations into visible controls. In an LLM interface, users can express higher-level goals while the system infers intermediate steps. This can reduce interface complexity, but it also moves important decisions into an opaque inference process. The resulting interaction is more flexible but also less deterministic.

**Table 1. Evolution of NLP technologies and their HCI implications**

| Stage                        | Typical period | Core NLP approach                           | HCI characteristics                              | Main limitation                                  |
|------------------------------|----------------|---------------------------------------------|--------------------------------------------------|--------------------------------------------------|
| Rule-based conversation      | 1960s–1990s    | Patterns, scripts, keywords                 | Simple text dialogue; fixed responses            | Brittle; little semantic understanding           |
| Statistical NLP              | 1990s–2010s    | Probabilistic language and speech models    | Search, dictation, intent classification         | Domain dependence; limited context               |
| Neural NLP                   | 2010s–2017     | Embeddings, RNN/LSTM, seq2seq               | Improved translation and generation              | Long-context and scalability limits              |
| Transformer NLP              | 2017–2022      | Self-attention; pre-trained representations | Contextual search, QA, flexible generation       | Large resource needs; opaque behaviour           |
| Generative conversational AI | 2022–2026      | Instruction-tuned LLMs, retrieval, tools    | Open-ended dialogue, copilots, multimodal agents | Hallucination, overreliance, privacy and control |

### A Human-Centered AI-NLP Interaction Architecture

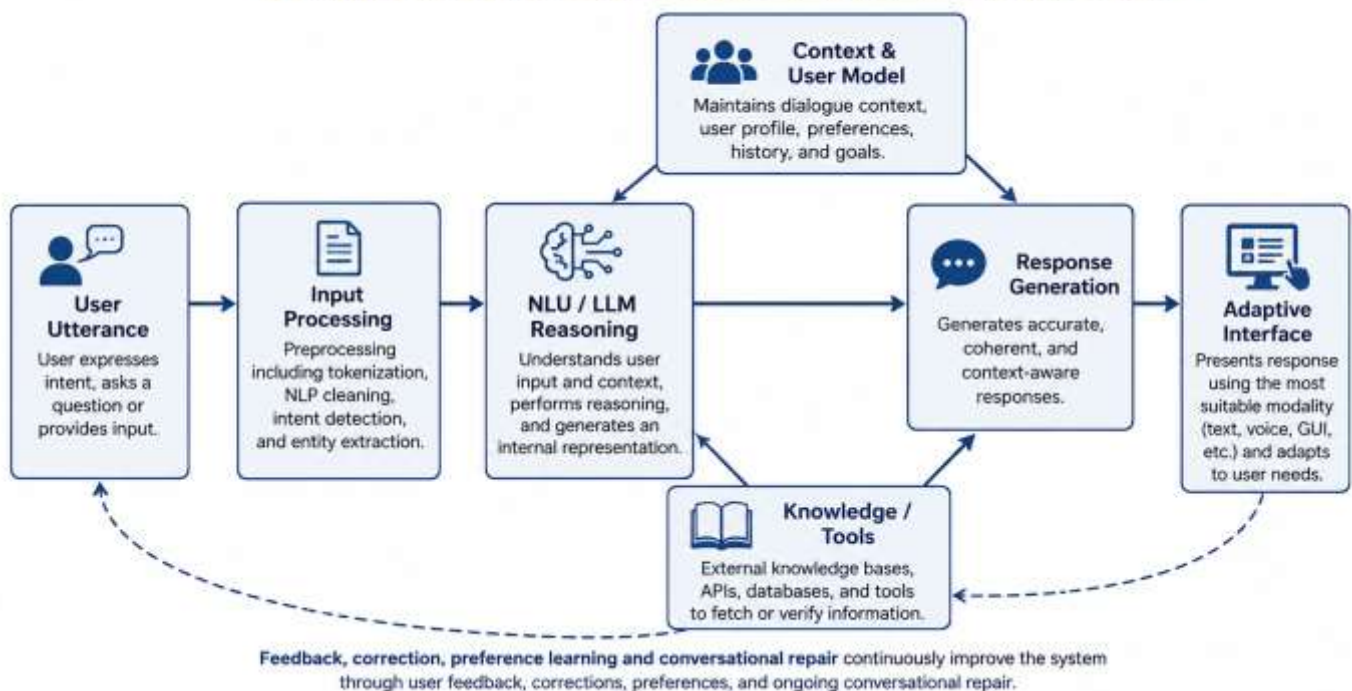


**Figure 2. Human-centered architecture for an AI-driven NLP interaction system.**

### 3. Literature Review

Research on conversational HCI consistently shows that users judge language interfaces through a combination of functional effectiveness and social expectations. Porcheron et al. (2018), studying voice interfaces in everyday settings, showed that interaction with voice assistants becomes embedded in ongoing social activity rather than occurring as an isolated user-device exchange. Their work highlights a central design issue: conversational systems must account for timing, turn-taking, accountability, and the surrounding activity. Clark et al. (2019) similarly found that users distinguish between social and functional qualities of conversation, suggesting that “human-like” behaviour is not always the primary goal; task appropriateness and common ground may matter more.

Usability research has attempted to formalize these conversational qualities. Borsci et al. (2022) developed the Chatbot Usability Scale to measure dimensions not adequately captured by traditional usability instruments, including conversational efficiency, accessibility, functionality, and interaction quality. Subsequent multilanguage validation produced a shorter BUS-11 instrument and showed that age can influence satisfaction ratings (Borsci et al., 2023). These findings are important for AI-driven NLP because they demonstrate that a fluent response alone is not an adequate usability measure. Conversational systems must be evaluated as interactive experiences, not only as language-generation engines.

Timing and social cues also influence user perception. In a laboratory experiment with 202 participants, Gnewuch et al. (2022) found that delayed chatbot responses had different effects for novice and experienced users. A delay could increase social presence for novices by resembling human response behaviour, but produce a negative effect for experienced users who expected chatbots to respond instantly. This finding illustrates that expectations are learned and user-specific. As natural-language interfaces become more common, designers cannot assume that a single conversational style will be optimal across users or contexts.

Trust is one of the most persistent themes in the literature. Følstad et al. (2018) identified interpretation quality, advice quality, human-likeness, self-presentation, professional appearance, privacy, security, and brand context as important trust factors in customer-service chatbots. The rise of LLMs complicates trust because model outputs can be highly fluent even when unsupported or incorrect. Bender et al. (2021) warned that large language models can reproduce biases and encourage users to over-attribute understanding to statistically generated language. Chiesurin et al. (2023) further showed that dialogue behaviours such as lexical alignment can increase perceived trust even when responses are not faithful to evidence, highlighting the danger of equating conversational polish with reliability.

Prompting introduces another usability layer. Zamfirescu-Pereira et al. (2023) observed non-AI experts trying to design LLM prompts and found that natural language does not automatically make system specification easy. Users struggled to build correct mental models of model behaviour and often used ad hoc trial and error. This means that LLM interfaces can lower some learning barriers while creating new ones: users may no longer need formal commands, but they must learn how to express constraints, provide context, judge output quality, and recover from model failures.

Recent empirical studies reinforce this point. Kim et al. (2024) collected 511 instances of dissatisfaction from 107 ChatGPT users and found that failure to grasp user intent was the most common source of dissatisfaction, while accuracy problems were rated as especially severe. Many attempts to resolve these failures were unsuccessful. Qian and Wexler (2024), studying 76 software engineers, found that conversational AI affected performance, efficiency, satisfaction, and trust differently depending on expertise and task type; they also reported evidence of automation complacency and increasing reliance. Manzoor et al. (2024), in a user study of ChatGPT as a conversational recommender system, found that generative dialogue can support user-centered recommendation experiences, but the value of the system depends on meaningful, personalized explanations and recommendations rather than generative capability alone.

The influence of AI-driven NLP also extends beyond text chat. Kim, Lee, and Mutlu (2024) compared an LLM-powered social robot with text- and voice-based agents in a user study of 32 participants. LLM-enabled interaction improved connection-building and deliberation, yet participants also expected richer non-verbal behaviour and reported weaknesses in logical communication and potential anxiety. Such findings show that language capability raises expectations for the entire interface. When an agent speaks or writes naturally, users may expect comparable competence in reasoning, perception, memory, embodiment, and social behaviour.

**Table 2. Representative empirical studies on conversational and AI-driven HCI**

| Study                            | Context / sample                             | Key HCI finding                                                                                      | Relevance to AI-NLP interfaces                                                       |
|----------------------------------|----------------------------------------------|------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| Porcheron et al. (2018)          | Month-long Amazon Echo deployments           | Voice-interface use is embedded in ongoing social interaction and turn-taking.                       | Conversational design must fit real social contexts, not isolated commands.          |
| Følstad et al. (2018)            | 13 chatbot users; interviews                 | Trust depends on interpretation/advice quality, presentation, privacy, security and service context. | Trust is multi-dimensional and contextual.                                           |
| Amershi et al. (2019)            | 49 design practitioners; 20 AI products      | Validated 18 guidelines for human-AI interaction.                                                    | AI interfaces need expectation setting, correction, feedback and graceful failure.   |
| Clark et al. (2019)              | Semi-structured interviews                   | Users distinguish social and functional conversation; common ground and context matter.              | Human-likeness should not replace task appropriateness.                              |
| Borsci et al. (2022)             | Four studies; 141 participants               | Developed chatbot usability measures beyond generic usability scales.                                | Conversational efficiency and interaction quality require dedicated evaluation.      |
| Gnewuch et al. (2022)            | Lab experiment; N=202                        | Response delays affect novices and experienced users differently.                                    | User expectations moderate social presence and usage intentions.                     |
| Zamfirescu-Pereira et al. (2023) | Non-AI experts designing prompts             | Natural-language prompting still requires mental models and iterative learning.                      | Language lowers syntax barriers but creates new interaction literacy needs.          |
| Kim et al. (2024)                | 107 users; 511 dissatisfaction cases         | Intent misunderstanding was frequent; accuracy dissatisfaction was severe; many repairs failed.      | Conversational repair and output verification remain core design problems.           |
| Qian & Wexler (2024)             | 76 software engineers                        | Productivity and trust effects vary by expertise/task; automation complacency observed.              | Benefits depend on task structure and user capability.                               |
| Kim, Lee, & Mutlu (2024)         | 32 participants; robot/text/voice comparison | LLM robots improved connection/deliberation but raised expectations and anxiety.                     | Better language capability changes expectations for the whole interface.             |
| Manzoor et al. (2024)            | Online study; N=190                          | Users evaluated LLM-generated recommendations and explanations positively in many scenarios.         | Generative dialogue can support personalization when explanations remain meaningful. |

#### 4. Methodology

This paper adopts an integrative narrative review methodology. The purpose is not to estimate a single pooled effect size, but to connect technical developments in NLP with empirical findings from HCI, conversational user interfaces, human–AI collaboration, and human–robot interaction. Literature was selected purposively from peer-reviewed or primary research sources that directly addressed one or more of the following: (a) enabling NLP architectures; (b) conversational usability and user experience; (c) trust and social perception; (d) user adaptation, prompting, and task performance; and (e) risks such as bias, opacity, and overreliance.

The synthesis emphasizes work from 2017 onward because the Transformer architecture represents a major technical transition in modern NLP, while also retaining selected earlier conversational-HCI studies where they illuminate enduring interaction problems. Sources include major outlets and proceedings in NLP and HCI, including NeurIPS, ACL/NAACL, ACM CHI, ACM IUI, ACM UMAP, HRI, Personal and Ubiquitous Computing, and Business & Information Systems Engineering. The review is interpretive rather than systematic: studies were compared for recurring mechanisms and outcomes, and findings were grouped into interaction naturalness, efficiency and productivity, personalization, accessibility, trust and transparency, user control, and safety.

To reduce the risk of technological determinism, the review treats NLP performance and HCI quality as related but distinct constructs. A technically stronger model may still produce a poor interaction if the interface creates inappropriate expectations, hides uncertainty, prevents correction, or encourages excessive reliance. Conversely, a more constrained system may produce a better experience in high-risk or highly structured tasks because its limits are clear. This distinction guides the conceptual framework proposed later in the paper.

## **5. Synthesis of Findings and Discussion**

The first major impact of AI-driven NLP is increased interaction naturalness. Users can formulate requests in their own words, ask follow-up questions, revise instructions, and shift between subtasks without navigating a rigid menu structure. This expands the “expressive bandwidth” of the interface: a single conversational surface can potentially expose many functions. In practical terms, language can act as a universal access layer over search, writing tools, analytics, customer service, knowledge systems, and software features. However, naturalness should not be confused with human imitation. The evidence reviewed here suggests that users value relevance, clarity, efficiency, and appropriate expectations at least as much as anthropomorphic behaviour (Clark et al., 2019; Gnewuch et al., 2022).

Second, NLP interfaces can improve efficiency by letting users state goals rather than execute every intermediate command. Generative assistants can summarize, transform, draft, explain, or search through information conversationally. Qian and Wexler (2024) show, however, that productivity effects are conditional: expertise and question type influence whether AI assistance helps. This is consistent with a broader HCI principle that automation changes work rather than simply removing it. Users must evaluate outputs, provide missing context, detect errors, and decide when to accept or override the system. When these oversight costs are ignored, apparent efficiency can be offset by verification burden or downstream correction.

Third, AI-driven NLP enables stronger personalization and context sensitivity. Unlike traditional task-oriented chatbots that depend on explicit slots, contemporary systems can use conversation history, user preferences, retrieved documents, and contextual signals to tailor responses. Personalization can reduce repetitive explanation and support continuity across multi-step tasks. Yet persistent context also creates privacy and boundary concerns. Designers must determine what is stored, what is inferred, how long memory persists, and how users can inspect or delete it. Personalization should therefore be paired with visible control rather than implemented as invisible behavioural inference.

Fourth, natural-language interaction creates important accessibility opportunities. It can reduce dependence on precise motor control, complex navigation, or application-specific literacy; voice interaction can support hands-busy or eyes-busy contexts; and multilingual language models can potentially broaden access to information and services. Borsci et al. (2023), however, show that usability evaluations can vary across age groups, and conversational systems often perform unevenly across languages, accents, dialects, and cultural conventions. Accessibility is therefore not an automatic property of NLP. It is an outcome that requires inclusive training data, multilingual evaluation, alternative modalities, adjustable interaction pace, and fallback mechanisms.

Fifth, AI-driven NLP transforms the social and psychological character of HCI. Language invites users to attribute agency, competence, personality, and even intention to a system. This can increase engagement but also produces overtrust. Bender et al. (2021) emphasize that linguistic fluency does not establish human-like understanding, while Chiesurin et al. (2023) demonstrate that conversational alignment can make unfaithful answers appear more trustworthy. The design challenge is therefore calibrated trust: users should rely on the system when it is appropriate and remain appropriately skeptical when uncertainty or risk is high.

Accuracy and transparency are central to calibrated trust. The dissatisfaction data reported by Kim et al. (2024) indicate that intent misunderstanding and incorrect information remain major pain points even in highly capable chat-based LLMs. A human-centered NLP interface should therefore support conversational repair. Users need

lightweight ways to say that the system misunderstood them, request evidence, ask what information was used, constrain the task, compare alternatives, or roll back an action. Amershi et al. (2019) similarly emphasize support for efficient correction and graceful failure in AI interfaces.

A further implication is that natural language should often complement rather than replace graphical interaction. Language is powerful for expressing goals, ambiguity, and open-ended requests, but GUIs remain superior for overview, comparison, precision, state visibility, and direct manipulation. A high-quality future interface may combine both: users describe a goal in language, while the system exposes assumptions, options, intermediate states, and consequences visually. Such hybrid interfaces can reduce the “black box” feeling of conversational AI and give users more opportunities to verify what the system understood before consequential actions occur.

Finally, the review suggests that evaluation methods must evolve. Model-centered metrics such as accuracy, F1 score, BLEU, or benchmark performance are insufficient for interactive systems. Evaluation should include task completion, time, cognitive effort, perceived control, trust calibration, error recovery, accessibility, satisfaction, and the quality of the user’s mental model. Instruments such as the BUS-11 offer one route for conversational usability, but LLM-based systems also require measures of verification burden, prompt repair, reliance, and harm.

**Table 3. Major HCI dimensions influenced by AI-driven NLP**

| HCI dimension   | Potential benefit of AI-driven NLP                              | Primary risk                                                 | Design response                                                               |
|-----------------|-----------------------------------------------------------------|--------------------------------------------------------------|-------------------------------------------------------------------------------|
| Naturalness     | Users express goals in everyday language.                       | Anthropomorphic over-attribution.                            | Use natural language while clearly signalling system identity and limits.     |
| Efficiency      | Fewer navigation steps; rapid transformation and synthesis.     | Verification burden and hidden errors.                       | Show intermediate state, sources, and editable outputs.                       |
| Learnability    | Reduces need to memorize commands or menus.                     | Prompting skill becomes a new hidden requirement.            | Offer examples, structured prompts, and clarification support.                |
| Personalization | Responses can adapt to history, preferences and context.        | Invisible profiling and privacy loss.                        | Provide memory controls, consent and inspectable personalization.             |
| Accessibility   | Voice and language can reduce motor/visual interaction demands. | Uneven performance across accents, languages and age groups. | Test inclusively; provide multi-modal alternatives and fallback paths.        |
| Trust           | Conversational explanation may improve confidence.              | Fluency can increase overtrust in incorrect output.          | Calibrate trust through evidence, uncertainty cues and correction mechanisms. |
| Collaboration   | Users can iteratively co-create and refine outputs.             | Responsibility becomes ambiguous.                            | Make action ownership, approval and reversibility explicit.                   |

**Table 4. Comparison of major interaction paradigms**

| Feature       | Traditional GUI               | Rule-based chatbot           | LLM-based conversational interface    |
|---------------|-------------------------------|------------------------------|---------------------------------------|
| User input    | Clicks, menus, forms          | Restricted phrases / intents | Open-ended natural language           |
| Task coverage | Explicitly designed functions | Narrow predefined domain     | Broad and dynamically specified tasks |

| Feature          | Traditional GUI                          | Rule-based chatbot             | LLM-based conversational interface                           |
|------------------|------------------------------------------|--------------------------------|--------------------------------------------------------------|
| Predictability   | Usually high                             | Moderate/high within domain    | Variable; probabilistic                                      |
| Context handling | Visible application state                | Limited slots/session state    | Potentially long conversational and retrieved context        |
| Learnability     | Requires navigation knowledge            | Requires discoverable phrasing | Low syntax barrier, but prompting/verification skills matter |
| Error mode       | Wrong action often visible               | Fallback or intent failure     | Plausible but incorrect or unsupported generation            |
| Best use         | Precision, overview, direct manipulation | Structured repetitive service  | Exploration, synthesis, drafting, flexible assistance        |
| Key HCI need     | Discoverability and consistency          | Intent recovery                | Trust calibration, evidence, control and repair              |

### Balancing Capability with Human Control in NLP-Based HCI

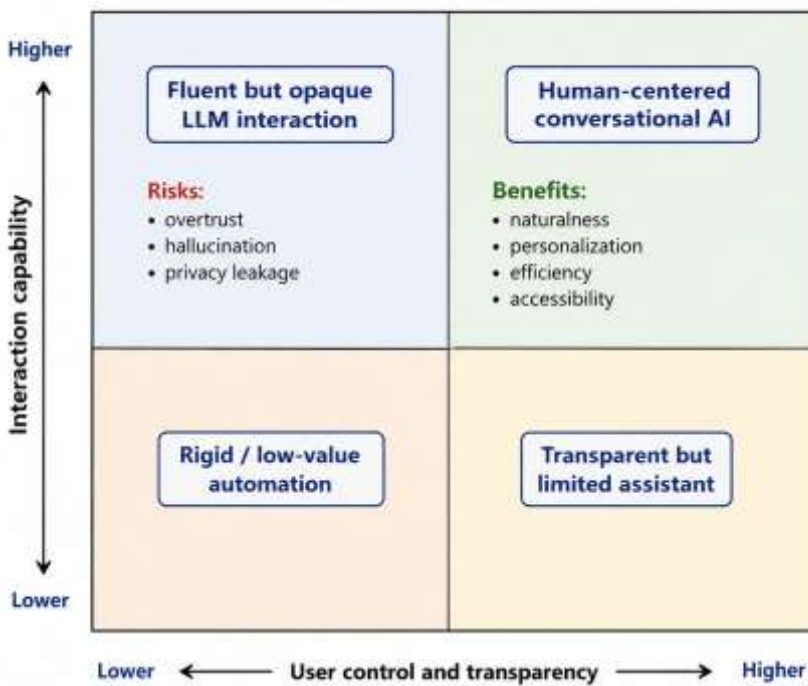


Figure 3. Capability–control matrix for evaluating benefits and risks of NLP-based interaction.

### 6. Proposed Human-Centered Framework for AI-NLP Interaction

The proposed framework conceptualizes the HCI impact of AI-driven NLP as a relationship between language capabilities, interaction quality, and user outcomes. Four capability clusters are central: natural-language understanding, context awareness and memory, personalization and adaptation, and multimodal accessibility. These capabilities influence an intermediate construct—AI-NLP interaction quality—which captures whether the system transforms technical language competence into a usable exchange. High interaction quality requires coherent turn-taking, appropriate response length, clear task progress, stable context, repair mechanisms, and alignment with the user’s actual goal.

Interaction quality then affects four broad HCI outcomes: usability and efficiency, engagement and social presence, task performance and human–AI collaboration, and inclusion/accessibility. Crucially, the strength and direction of these effects are moderated by trust, accuracy, transparency, and user expertise. The same fluent assistant may help

a novice on an exploratory task but mislead an expert on a precision-critical task if evidence and uncertainty are hidden. Boundary conditions—including privacy, bias, safety, and cultural or linguistic fit—can also limit whether a technically capable system delivers equitable benefits.

The framework therefore rejects a simple assumption that more capable language models necessarily produce better HCI. Instead, it proposes a conditional relationship: capability improves interaction only when the interface helps users understand what the system can do, maintain control over the interaction, detect and recover from errors, and calibrate reliance to the quality of the underlying evidence.

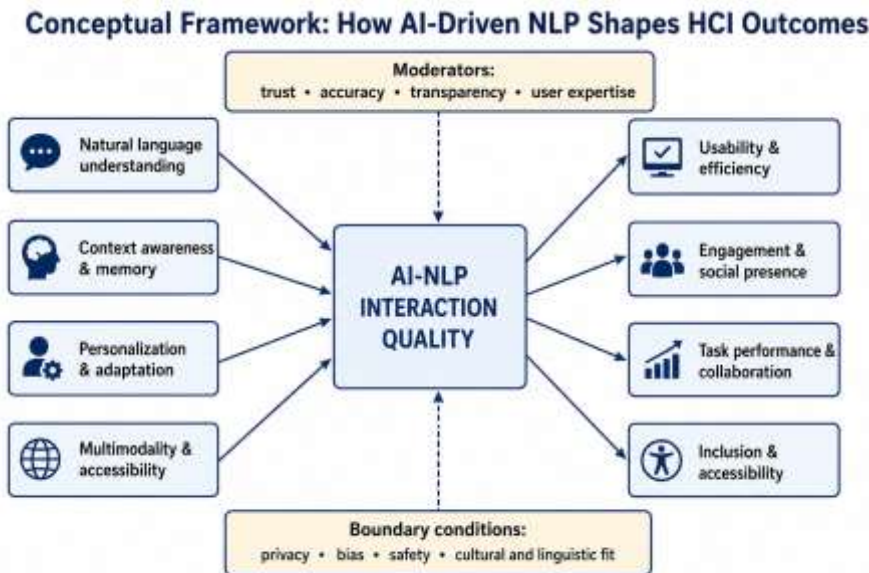


**Figure 4. Proposed conceptual framework linking AI-NLP capabilities to HCI outcomes.**

## 7. Design and Managerial Implications

For designers, the first implication is to design for intent negotiation rather than one-shot intent recognition. When a request is ambiguous, the system should ask targeted clarifying questions or present interpretations instead of confidently committing to one. Second, interfaces should provide visible progress and state. Users need to know what information the system is using, which step it is performing, and whether an answer is generated from memory, retrieved evidence, computation, or external tools.

Third, conversational systems should make correction cheap. A user should be able to refine constraints, reject an assumption, edit context, or revert an action without restarting the interaction. Fourth, designers should calibrate anthropomorphism. Human-like language can improve engagement, but personality cues should not obscure the system's artificial nature or its limitations. Fifth, high-stakes applications should expose sources, confidence-relevant information, and escalation routes to human support where appropriate.

For organizations, AI-driven NLP should be treated as an interaction redesign project rather than a model deployment project. Successful adoption requires usability testing, content governance, privacy controls, monitoring of failure patterns, domain-specific safeguards, and training for users who must evaluate AI outputs. The evidence on prompt difficulty and expertise effects indicates that organizations should not assume employees or customers will intuitively know how to collaborate with LLM systems (Zamfirescu-Pereira et al., 2023; Qian & Wexler, 2024). Evaluation should also be segmented by user group and task. Gnewuch et al. (2022) show that even response timing can affect novices and experienced users differently. Similar variation is likely for explanation depth, response length, interface density, and proactive suggestions. Adaptive systems can address such differences, but adaptation must remain inspectable and reversible.

Table 5. Design priorities and future research agenda

| Priority | Design / research recommendation                      | Rationale                                                  | Example evaluation indicator                         |
|----------|-------------------------------------------------------|------------------------------------------------------------|------------------------------------------------------|
| 1        | Design for conversational repair                      | Misunderstanding is inevitable in open-ended interaction.  | Repair success rate; turns to recovery               |
| 2        | Calibrate trust, not maximum trust                    | Fluent output can be wrong yet persuasive.                 | Appropriate reliance; error detection                |
| 3        | Combine language with visible state                   | Conversation can hide assumptions and progress.            | State comprehension; task reversibility              |
| 4        | Segment by user expertise and context                 | Novices and experienced users form different expectations. | Task success by expertise group                      |
| 5        | Evaluate inclusivity across language, age and ability | Average usability can conceal exclusion.                   | Performance and BUS/UX measures by subgroup          |
| 6        | Expose provenance and tool use                        | Users need to know where answers and actions come from.    | Source inspection; evidence use                      |
| 7        | Measure long-term effects                             | Trust, reliance and skill change with repeated use.        | Longitudinal reliance, skill retention, satisfaction |
| 8        | Define privacy and memory boundaries                  | Contextual personalization can become surveillance.        | User understanding/control of stored context         |

### Continuous Design and Evaluation Cycle for Conversational HCI



Figure 5. Continuous human-centered design and evaluation cycle for conversational AI.

## 8. Limitations and Future Research

This paper is an integrative narrative review and therefore does not claim exhaustive coverage of all NLP-HCI research or provide a meta-analytic estimate of effect sizes. The field is also changing rapidly, which means specific model capabilities and interface conventions can evolve faster than peer-reviewed evidence. In addition, many current studies are conducted with relatively short interactions, whereas long-term use may produce different patterns of trust, skill development, dependency, and expectation.

Future research should prioritize longitudinal studies of human–LLM collaboration, cross-cultural and multilingual evaluation, accessibility for older adults and users with disabilities, calibrated trust measures, and comparative research on hybrid GUI-plus-language interfaces. More work is also needed on how users form mental models of LLM uncertainty, how conversational repair should be designed, and how interfaces can communicate provenance without overwhelming users. In high-stakes domains, research should examine when natural-language interfaces should defer, refuse, or transfer control to humans rather than maximizing conversational continuity.

## 9. Conclusion

AI-driven natural language processing is reshaping HCI by moving interaction from predefined commands toward flexible, goal-oriented conversation. Transformer-based models and instruction-tuned LLMs have made it possible for users to query, instruct, revise, and collaborate with systems through ordinary language across an unprecedented range of tasks. The resulting interfaces can lower learning barriers, support personalization, expand accessibility, and increase the expressive power of human–computer interaction.

The same qualities that make conversational AI attractive also create new risks. Fluent language can mask uncertainty, encourage overtrust, amplify bias, and make system failures harder to detect. Empirical HCI research shows that user experience depends not only on linguistic quality but on accuracy, context, timing, expertise, transparency, privacy, and the availability of correction mechanisms. The most productive direction for future HCI is therefore not to make computers merely sound more human. It is to combine natural language with clear system state, evidence, user control, inclusive design, and continuous evaluation. In that form, AI-driven NLP can become a powerful interface layer that supports rather than displaces human judgment.

## 10. References

### References within Main Content of the Research Paper

- [1] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human–AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3290605.3300233>
- [2] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- [3] Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The Chatbot Usability Scale: The design and pilot of a usability scale for interaction with AI-based conversational agents. *Personal and Ubiquitous Computing*, 26(1), 95–119. <https://doi.org/10.1007/s00779-021-01582-9>
- [4] Borsci, S., Schmettow, M., Malizia, A., Chamberlain, A., & van der Velde, F. (2023). A confirmatory factorial analysis of the Chatbot Usability Scale: A multilanguage validation. *Personal and Ubiquitous Computing*, 27(2), 317–330. <https://doi.org/10.1007/s00779-022-01690-0>
- [5] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- [6] Chiesurin, S., Dimakopoulos, D., Sobrevilla Cabezudo, M. A., Eshghi, A., Papaioannou, I., Rieser, V., & Konstas, I. (2023). The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering. *Findings of the Association for Computational Linguistics: ACL 2023*, 947–959.
- [7] Clark, L., Pantidi, N., Cooney, O., Doyle, P. R., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., & Cowan, B. R. (2019). What makes a good conversation? Challenges in designing truly conversational agents. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Article 475. <https://doi.org/10.1145/3290605.3300705>

- [8] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [9] Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. In S. S. Bodrunova (Ed.), *Internet Science (LNCS 11193)*, pp. 194–208). Springer. [https://doi.org/10.1007/978-3-030-01437-7\\_16](https://doi.org/10.1007/978-3-030-01437-7_16)
- [10] Fossa, F., & Sucameli, I. (2022). Gender bias and conversational agents: An ethical perspective on social robotics. *Science and Engineering Ethics*, 28, 23. <https://doi.org/10.1007/s11948-022-00376-3>
- [11] Gnewuch, U., Morana, S., Adam, M. T. P., & Maedche, A. (2022). Opposing effects of response time in human–chatbot interaction: The moderating role of prior experience. *Business & Information Systems Engineering*, 64(6), 773–791. <https://doi.org/10.1007/s12599-022-00755-x>
- [12] Kim, C. Y., Lee, C. P., & Mutlu, B. (2024). Understanding large-language model (LLM)-powered human-robot interaction. *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 371–380. <https://doi.org/10.1145/3610977.3634966>
- [13] Kim, Y., Lee, J., Kim, S., Park, J., & Kim, J. (2024). Understanding users’ dissatisfaction with ChatGPT responses: Types, resolving tactics, and the effect of knowledge level. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 385–404. <https://doi.org/10.1145/3640543.3645148>
- [14] Manzoor, A., Ziegler, S. C., Pirker Garcia, K. M., & Jannach, D. (2024). ChatGPT as a conversational recommender system: A user-centric analysis. *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, 267–272. <https://doi.org/10.1145/3627043.3659574>
- [15] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- [16] Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. <https://doi.org/10.1145/3173574.3174214>
- [17] Qian, C., & Wexler, J. (2024). Take it, leave it, or fix it: Measuring productivity and trust in human–AI collaboration. *Proceedings of the 29th International Conference on Intelligent User Interfaces*.
- [18] Schouten, D. G. M., Deneka, A. A., Theune, M., Neerinx, M. A., & Cremers, A. H. M. (2023). An embodied conversational agent coach to support societal participation learning by low-literate users. *Universal Access in the Information Society*, 22, 1215–1241. <https://doi.org/10.1007/s10209-021-00865-5>
- [19] See, A., Roller, S., Kiela, D., & Weston, J. (2019). What makes a good conversation? How controllable attributes affect human judgments. *Proceedings of NAACL-HLT 2019*, 1702–1723.
- [20] Teixeira, M. S., & Dragoni, M. (2022). A review of plan-based approaches for dialogue management. *Cognitive Computation*, 14, 1019–1038. <https://doi.org/10.1007/s12559-022-09996-0>
- [21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- [22] Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can’t prompt: How non-AI experts try (and fail) to design LLM prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Article 437. <https://doi.org/10.1145/3544548.3581388>