

HIPA: A HUMAN INFORMATION PROCESSING ARCHITECTURE FOR LLM RESPONSES IN ERGONOMIC DESIGN

Youngjun Park¹, Jaekyu Park² and Youngjae Im^{1,3,*}

¹Department of Ergonomics Engineering
Dong Eui University
Busan, Korea

²Department of Bigdata and Management Engineering
Namseoul University
Cheonan, Korea

³The Research Institute of UX
Dong Eui University
Busan, Korea

*Corresponding author's e-mail: ergoim@deu.ac.kr

Large language models (LLMs) are increasingly used for UI/UX evaluation and design decision support; however, their responses are often generated as implicit and unstructured outputs, making cognitive processes and judgment rationale difficult to trace. Moreover, the quality of LLM-based judgments may vary depending on users' prompting proficiency, limiting the consistency of AI-assisted evaluation. This study proposes HIPA (Human Information Processing Architecture), a cognitive response architecture that structures LLM outputs based on the human information processing (HIP) model and investigates the manifestation of human cognitive stages in conventional LLM responses. HIPA maps four HIP stages—perception, working memory, decision making, and response execution—into a two-stage response structure consisting of information provision and judgment support through knowledge-grounded prompting with GPT-4o. A formative study was then conducted with seven participants of differing prompting proficiency, who assessed the same mobile login interface using a general-purpose LLM without HIPA. Responses were coded according to HIP-stage manifestation using a three-point scale. The results showed that perception and response execution were consistently observed, whereas working memory and decision-making stages varied depending on prompt systematicity. These findings indicate that general LLM responses do not inherently provide cognitively structured judgment support. The study therefore tests the structural premise on which HIPA is built, not HIPA's own outputs; a direct comparison of HIPA and general-LLM responses is identified as the next step. HIPA provides a human-centered framework for developing AI-assisted UI/UX evaluation systems based on human cognitive principles.

Keywords: Human Information Processing; Large Language Models; UI/UX Evaluation; Decision Support

(Received on July 27, 2026; Accepted on September 23, 2026)

1. INTRODUCTION

The rapid advancement of large language models (LLMs) has led UI/UX design practitioners and researchers to routinely rely on generative AI tools such as ChatGPT for usability evaluation, comparison of design alternatives, and guideline verification. Because professional advice can be obtained instantly from a natural-language query alone, AI has quickly become established as a collaborative partner that reduces users' cognitive burden (Xu & Dainoff, 2023; Duan, Warner, *et al.*, 2024). As AI adoption becomes more widespread in practice, however, the question of how far its responses can be trusted and incorporated into decision-making inevitably arises. General-purpose LLMs are adept at producing fluent and plausible conclusions, but it is difficult for users to trace and verify the evidence and reasoning process behind those conclusions (Rawte *et al.*, 2023; Bach *et al.*, 2024; Liao & Vaughan, 2023).

This lack of verifiability leads to risks that go beyond mere inconvenience. LLMs can undermine response reliability and user trust through hallucination—the plausible generation of factually incorrect content (Ji *et al.*, 2023)—and tend to rely on their internal training data to produce poorly grounded conclusions, particularly when users enter vague or unstructured prompts. A more serious problem is that users may mistake an AI's fluent phrasing for evidence of its validity, leading to overreliance, in which users accept and act on responses that are in fact incorrect (Klingbeil *et al.*, 2024). The less visible the reasoning process is, the harder it becomes for users to self-check the appropriateness of a conclusion, increasing the risk that unverified judgments are carried directly into practical decision-making.

This problem is especially severe in the domain of UI/UX evaluation. UI/UX evaluation is not a task with a single correct answer; rather, it is a judgment task that requires weighing the severity of issues against multiple evaluation criteria and prioritizing among them (Kim & Im, 2024; Oh *et al.*, 2025; Duan, Cheng, *et al.*, 2024). Consequently, if an AI in this domain provides only a conclusion, its utility is limited; only when the evaluation criteria on which the conclusion is based and the comparative reasoning behind it are presented together can practitioners verify its validity and make responsible decisions. In other words, what this domain requires is not a more accurate answer but a response in which the grounds and process of judgment are made transparent.

Efforts to reflect human thought processes in design have a long tradition in the field of ergonomics. The Human Information Processing (HIP) model, which characterizes human information processing as a sequence of perception, working memory, decision making, and response execution, has been widely used as a core theoretical framework for interface design and task analysis (Wickens & Carswell, 2021; Wickens & Hollands, 2000). Related computational accounts such as ACT-R have been used to model user task performance (Ritter *et al.*, 2019), but, like the HIP model, they have been applied to the human side of the interaction rather than to the structure of a system's output. Cognitive load theory (Sweller, 1988), which addresses the capacity limits of working memory, underscores the importance of structuring information into a form users can process. In the UI/UX field, Nielsen's usability heuristics (Nielsen, 1994) have likewise been widely applied as a criterion for systematically diagnosing usability problems.

More recently, domain-specific generative AI systems that reflect the knowledge and task context of a particular domain have been proposed across various practical fields. Park and Lee (2025) implemented a product-information question-answering system that combines product manuals and standardized consultation data in a vector database using a Retrieval-Augmented Generation (RAG) architecture. In a qualitative evaluation with customer service agents, that system achieved a response accuracy of over 80%. Won (2025) developed a RAG-based domain-specific GPT to support ground control station (GCS) interface design for unmanned aerial vehicles, confirming that a model enhanced with domain knowledge provided more accurate and useful information than a general-purpose ChatGPT. That study also reported that AI usage needs differ markedly between experts, who prioritize reflecting real-world processes, and novices, who prioritize systematic learning. In this way, domain-specific approaches have continued to emerge as attempts to compensate for the limitations of general-purpose models.

In parallel, work on eliciting explicit reasoning from LLMs has proposed chain-of-thought prompting, which guides a response through a stepwise reasoning process; this line of research reports that making the reasoning structure explicit improves the quality of a model's judgments (Wei *et al.*, 2022; Wang *et al.*, 2023). Building on this, Yao *et al.* (2023) extended the single-path chain of thought into a tree structure, proposing a Tree-of-Thoughts approach that explores multiple reasoning paths and allows backtracking to earlier steps when necessary, improving performance on complex problem-solving tasks. A related line of work decomposes a task into chained prompts to make the interaction more transparent and controllable (Wu *et al.*, 2022).

However, these prior studies do not adequately address the problem this study aims to tackle, for two reasons. First, cognitive models and usability theories have mainly been applied to designing and evaluating human users or interfaces, and cases in which the reasoning structure of an AI response itself is explicitly designed according to a cognitive model and its manifestation verified remain rare. Second, approaches that attempt to improve response structure, such as chain-of-thought prompting, make securing that structure dependent on individual users' prompt-writing proficiency, and non-experts are known to struggle with prompt design (Zamfirescu-Pereira *et al.*, 2023; Liu *et al.*, 2023). In this case, less proficient users cannot reliably obtain the cognitive structure needed for judgment even when using the same model, resulting in a variability problem in which the cognitive structure of a response is governed by an individual user's prompting skill. Ultimately, the problem of guaranteeing, at the system level and independent of user proficiency, the cognitive structure required for judgment remains unresolved.

In response to the limitations of prior work, this study designs and implements HIPA (Human Information Processing Architecture), a UI/UX domain-specific assistant that directly adopts the Human Information Processing (HIP) model as a design criterion for AI response generation. HIPA automatically branches into a Stage 1 (information-provision) or Stage 2 (judgment-support) response structure depending on question type. For Stage 2 responses that require judgment, it provides an expert-style reasoning structure segmented into four core sections—[Interpretation], [Criteria], [Analysis], and [Recommendation]—together with an [Executive Summary], by default and independent of the user's prompt. The core aim of this study is not to improve the underlying response quality of the LLM itself, and it is therefore distinct from research on prompt-writing techniques alone. What HIPA implements is not response accuracy but the explicitness of response structure: whether, within the same model, the reasoning structure is left implicit or explicitly segmented. The effect of that structure on users is not measured in the present study, which establishes where the implicit structure falls short.

The contributions of this study are as follows. First, it presents a design principle intended to supply the cognitive structure required for judgment at the system level, without depending on users' prompting proficiency. This offers the practical value of consistent judgment support independent of proficiency variation. Second, by examining whether the cognitive stages of human information processing manifest in general LLM responses, it identifies where that structure is not reliably obtained. On this basis, it motivates the application of cognitive-engineering design principles to AI response design, and it presents the academic significance of structural transparency and cognitive auditability in

responses. The remainder of this paper is organized as follows: Section 2 describes the system design and development of HIPA; Section 3 describes system implementation and the user interface; Section 4 describes the validation method and analysis results; and Section 5 presents the discussion and conclusion.

2. SYSTEM DESIGN AND DEVELOPMENT

2.1. System Design Framework

HIPA, developed in this study, is a UI/UX domain-specific AI assistant built using the HIP model as its design criterion. The HIP model describes human cognitive processing as four stages—perception, working memory, decision making, and response execution (Wickens *et al.*, 2004)—and this study designed a response structure corresponding to each cognitive stage on the premise that the process by which a user receives, understands, and applies an AI response to judgment follows the same cognitive flow. Whereas general-purpose LLMs have an implicit response structure that returns an answer to a question as a single block of text, HIPA reflects the user’s cognitive processing flow by explicitly segmenting the response into structured sections.

The system consists of two layers: a user interface layer and an automated backend layer. The user interface is implemented as a single HTML file, and the backend is linked to OpenAI models through a Make.com webhook scenario, using GPT-4o for response generation and GPT-4o-mini for question classification. An incoming question has its type determined on the backend and is then automatically routed to either the Stage 1 information-provision or Stage 2 judgment-support response path. The overall system configuration and data flow are shown in Figure 1.

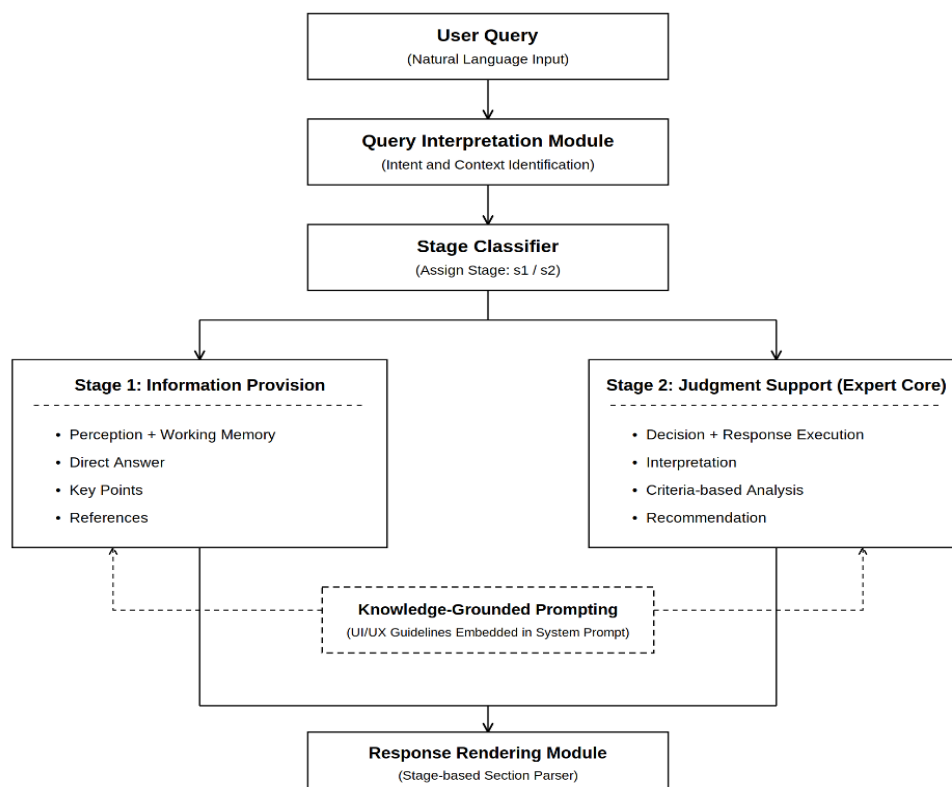


Figure 1. System Architecture and Data Flow of HIPA

The mapping between the four HIP stages and the HIPA response structure is shown in Table 1. The Stage 1 response structure applies to fact-checking or evidence-seeking questions that require perception- and working-memory-level processing, whereas the Stage 2 response structure applies to judgment- and decision-support questions that require decision-making- and response-execution-level processing.

Perception and Working Memory are grouped into Stage 1 because both concern the input side of information processing—accurately receiving and holding the relevant facts—whereas Decision Making and Response Execution are grouped into Stage 2 because both concern the central-processing and output side—forming a judgment from that information and converting it into action. This grouping follows the standard architecture of the HIP model, in which perception and working memory constitute the encoding/comprehension stages and decision making and response execution constitute the central-decision and output stages (Wickens *et al.*, 2022).

For each stage, the corresponding HIPA section addresses a specific cognitive burden with a design principle known to relieve it.

Perception. Because attention is a capacity-limited resource, the cost of locating the answer must be minimized before it competes with later-stage resources (Wickens *et al.*, 2022; Treisman & Gelade, 1980). Information seeking is governed by the cost of locating information relative to its expected value (Pirolli & Card, 1999), and readers treat initially mentioned content as thematic (Kieras, 1980); placing the answer at the point of lowest search cost follows, which [Answer] implements as a visually separated block at the top of the response.

Working memory. Capacity is sharply limited—classically estimated at seven plus or minus two chunks (Miller, 1956) and, on later estimates that control for rehearsal and chunking, closer to four (Cowan, 2001), and separating claims from their supporting evidence imposes an added split-attention cost (Chandler & Sweller, 1992). [Key Points] pre-chunks the response, and [Reference] co-locates the supporting evidence.

Decision making. Under cognitive load, people default to non-compensatory heuristics and lose track of their own criteria (Tversky & Kahneman, 1974; Payne *et al.*, 1993), and externalized criteria reduce overreliance when verifying them costs the user less than deferring to the system (Vasconcelos *et al.*, 2023). [Interpretation], [Criteria], and [Analysis] externalize exactly this.

Response execution. Understanding does not automatically convert into action (Norman, 1986); specifying the concrete action closes part of that gap (Gollwitzer, 1999), and decision support is most effective when an actionable recommendation is delivered at the point of decision (Kawamoto *et al.*, 2005). [Recommendation] and [Executive Summary] implement this at the point of decision.

Table 1. Mapping between HIP Model Stages and HIPA Response Structure

HIP Stage	Cognitive Process	Response Stage	Response Section
Perception	Stimulus detection Information recognition	Stage 1 (Information Provision)	[Answer]
Working Memory	Information interpretation Knowledge activation	Stage 1 (Information Provision)	[Key Points] [Reference]
Decision Making	Alternative evaluation Criteria application	Stage 2 (Judgment Support)	[Interpretation] [Criteria] [Analysis]
Response Execution	Optimal action selection Execution	Stage 2 (Judgment Support)	[Recommendation] [Executive Summary]

2.2. Stage-wise Response Algorithm

2.2.1. Stage 1: Information-Provision Response

The Stage 1 response structure applies to fact-checking or evidence-seeking questions and corresponds to the perception and working-memory processing levels of the HIP model. This structure consists of three sections: [Answer], [Key Points], and [Reference]. The [Answer] section provides a direct, concise, factual response to the question. The [Key Points] section itemizes the core content of the response to reduce the user's working-memory load, and the [Reference] section specifies the standards or guidelines—such as WCAG 2.1, Nielsen's 10 Heuristics, and Material Design—on which the response is grounded, thereby making the basis of the response inspectable, so that a user can verify the cited standard against its source. The design principle for this stage is conciseness of information and explicitness of evidence.

2.2.2. Stage 2: Judgment-Support Response

The Stage 2 response structure applies to questions requiring judgment and decision support and corresponds to the decision-making and response-execution processing levels of the HIP model. This structure consists of four core sections—[Interpretation], [Criteria], [Analysis], and [Recommendation]—together with an [Executive Summary]. The [Interpretation] section reinterprets the user's question from a UI/UX domain perspective to clarify the context of the problem, and the [Criteria] section presents the key considerations and related principles required for judgment. The [Analysis] section explicitly reveals the reasoning process by comparing the strengths and weaknesses of alternatives according to the established criteria, and the [Recommendation] section converts the analysis into concrete improvements for each issue, ordered by severity. The [Executive Summary] is an expert summary block that condenses the core rationale for judgment and the final recommendation; users can download the full structured response as a text file for use as material in practical decision-making. In the implementation, these sections are labeled [Question Interpretation] and [Decision Criteria]; the abbreviated forms [Interpretation] and [Criteria] are used throughout this paper. The core design principle for this stage is visibility of the reasoning process: by structuring the logical process leading to a judgment so that it can be traced and verified by the user, HIPA implements an explicit reasoning structure distinct from the single-block text responses of general-purpose LLMs.

2.3. Knowledge-Grounded Prompting Design

A representative approach for incorporating domain knowledge is vector-database-based Retrieval-Augmented Generation (RAG); RAG has also been widely adopted in practical systems, such as a system that analyzes discussion trends in DevOps-related online communities (Yoon & Kim, 2025) and a personalized web service that recommends travel destinations in real time based on user context (Jeong *et al.*, 2024), but the present system does not use this approach. Instead, it adopts Knowledge-Grounded Prompting, in which the system prompt specifies the response section structure and directs the model to ground each response in named UI/UX standards—such as WCAG 2.1, Nielsen’s 10 Heuristics, and Fitts’s Law—while explicitly prohibiting the citation of standards the model is not certain of. Domain expertise is thus enforced at the level of prompt design rather than through a separate knowledge base, which is advantageous for enforcing a consistent response structure. The system prompts for Stage 1 and Stage 2 each specify the response section structure and the standards to be cited, so that every response conforms to a predefined structure. The prompts are reproduced in full in the Appendix. A comparison of the system-prompt design principles for the two stages is shown in Table 2.

Table 2. Comparison of System Prompt Design Principles between Stage 1 and Stage 2

Category	Stage 1 — Information Provision	Stage 2 — Judgment Support
Response Goal	Provide accurate facts and evidence	Explicit structuring of the judgment process
Response Section	[Answer] [Key Points] [Reference]	[Interpretation] [Criteria] [Analysis] [Recommendation] [Executive Summary]
HIP Mapping	Perception + Working Memory	Decision Making + Response Execution
Domain Knowledge	Cite relevant standards and criteria	Present judgment rationale based on design principles
Design Principle	Conciseness · Evidence clarity	Reasoning visibility · Single recommendation
Output	Structured text response	Text response + Executive Summary (downloadable .txt)

Question classification is performed by the first OpenAI module in the Make.com scenario. This module analyzes the intent and type of the incoming question, determines whether it belongs to Stage 1 or Stage 2, and passes it to the router module. Requests to confirm objective facts, reference figures, or standard guidelines are classified as Stage 1, while questions that require evaluating a concrete interface, selecting a design direction, comparing alternatives, or deriving a recommendation are classified as Stage 2. When the classifier output matches neither category label, an error response is returned via a fallback path.

3. SYSTEM IMPLEMENTATION AND USER INTERFACE

3.1. User Interface Implementation

HIPA’s user interface is implemented as a single HTML file, adopting a layout composed of a sidebar and a main area. The sidebar permanently displays the brand wordmark, description cards for the Stage 1 and Stage 2 response modules, and system information, allowing users to check the difference between the two response structures at any point during a conversation. In its initial empty state, the main area provides four recommended UI/UX-related question chips to support the user’s first query, and it switches to a scrolling message area once the conversation begins. The input box at the bottom supports automatic height adjustment and provides Enter-to-send and Shift+Enter-for-newline shortcuts. The initial interface screen is shown in Figure 2.

3.2. Stage-wise Response Rendering

HIPA automatically determines its rendering method based on the stage field value in the backend webhook response. When a Stage 1 value is received, the Stage 1 section parser is invoked to render the [Answer], [Key Points], and [Reference] sections; when a Stage 2 value is received, the Stage 2 section parser and the Executive Summary block renderer are invoked together. This automatic branching allows users to receive a response structure appropriate to the question type without having to switch modes manually.

3.2.1. Stage 1 Response Screen

When a Stage 1 response is received, a Stage 1 information-provision stage tag is displayed at the top of the message area, followed in order by the [Answer], [Key Points], and [Reference] sections. Each section is visually distinguished

by a unique color label and a left-side vertical accent line. The actual rendered screen for a Stage 1 response is shown in Figure 3.

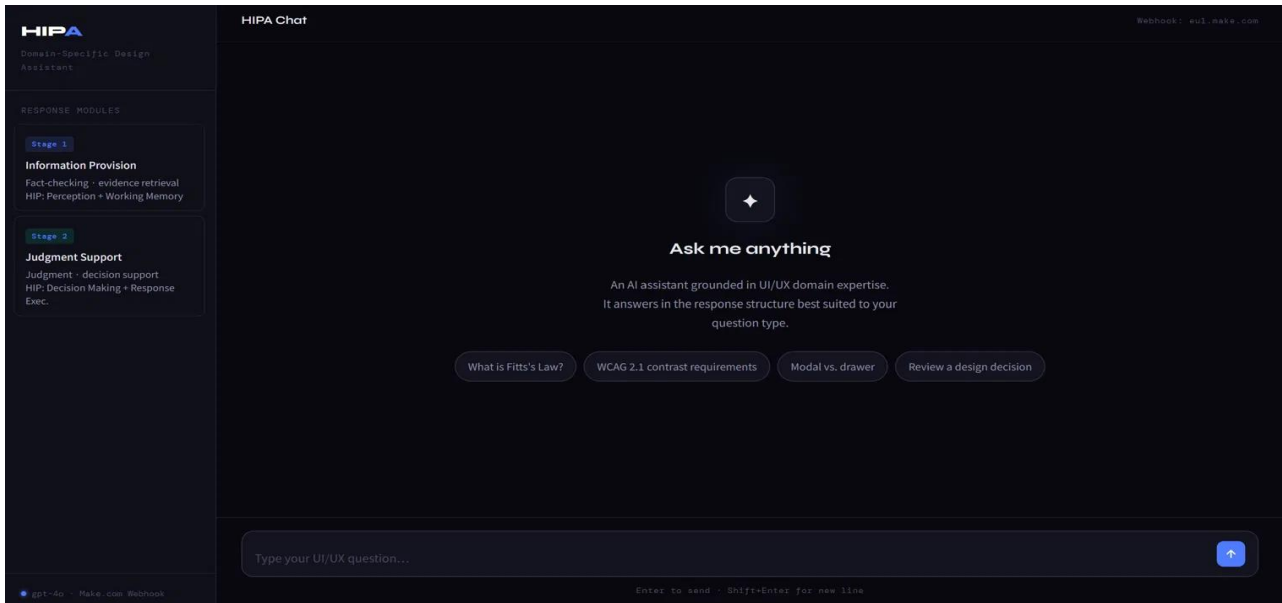


Figure 2. Initial Interface of HIPA

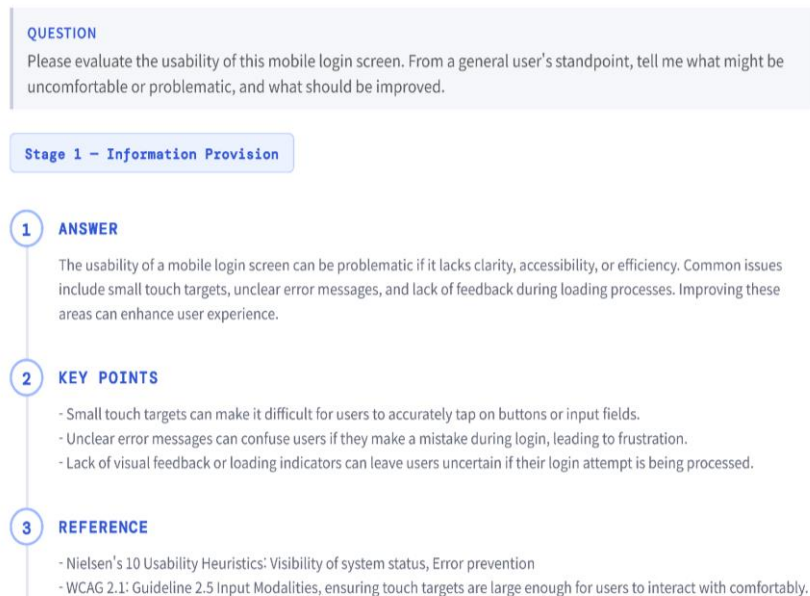


Figure 3. Stage 1 Response Rendering — Information Provision

3.2.2. Stage 2 Response Screen

When a Stage 2 response is received, the four sections—[Interpretation], [Criteria], [Analysis], and [Recommendation]—are rendered in order together with a Stage 2 judgment-support stage tag. The [Executive Summary] section is displayed separately at the bottom as a distinct highlighted block, below which an Executive Summary download button is provided; clicking it saves the content of all sections of that response as a structured text file for use as material in practical decision-making. Rather than being listed independently, the four sections are connected into a single line of reasoning, whose structure is shown in Figure 4. [Interpretation] parses the screen elements and the intent of the question to define the subsequent evaluation scope for judgment; based on this scope, [Criteria] selects the standards or heuristics to be applied, which become the judgment basis for [Analysis]. [Analysis] maps each issue to a criterion and assigns it a severity, yielding a severity-ranked set of issues, and [Recommendation] promotes the issue with the highest severity to the final recommendation.

This dependency between sections can also be confirmed as a single traceability thread on the actual response screen. The rendered screen for an actual Stage 2 response is shown in Figure 5, where numbered coral highlights and a connecting line on the right indicate the path along which a single issue is traced from [Interpretation] to [Executive Summary]. For example, the login-screen low-contrast issue shown in Figure 5 is carried, section by section, from [Interpretation] through to [Executive Summary] as a single traceable line of reasoning; the remaining issues are abridged to their titles and severity levels only, to keep the structure clear.

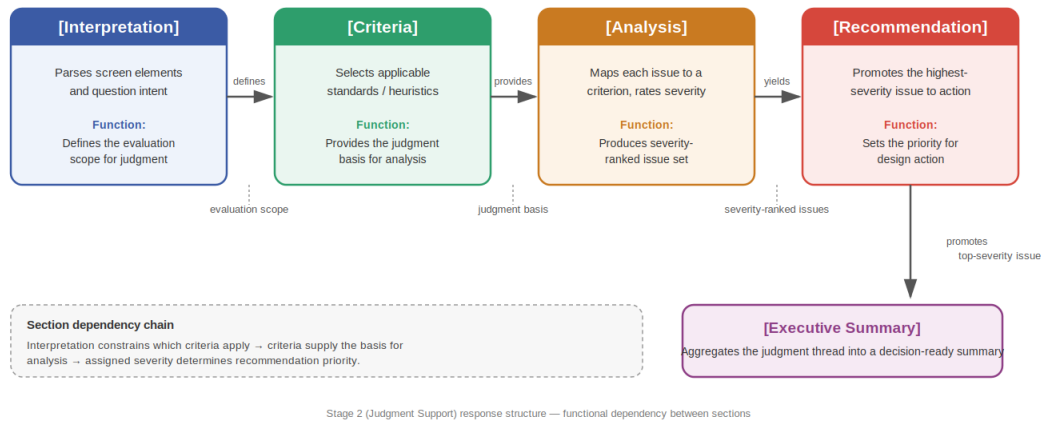


Figure 4. Stage 2 Response Structure Schema: Functional Dependency between Sections

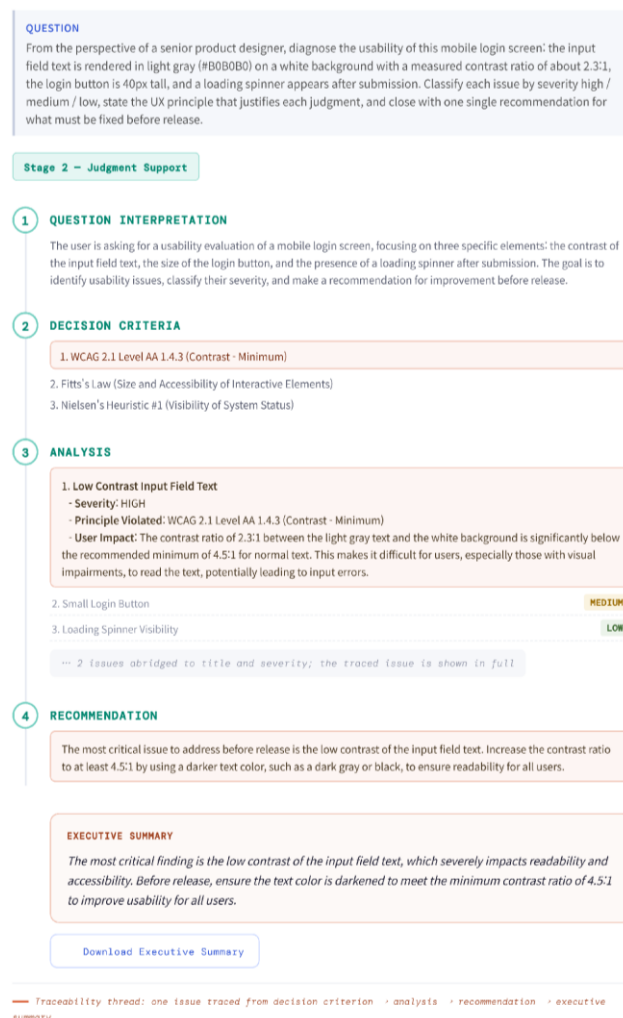


Figure 5. Stage 2 Response Capture with Traceability Thread Annotation: Question Interpretation → Decision Criteria (WCAG 2.1 1.4.3) → Analysis (Low Contrast Text, HIGH) → Recommendation / Executive Summary

3.3. Backend Automation Configuration

The backend consists of a Make.com webhook scenario connecting a total of eight modules. A user question is received by the webhook module via an HTTP POST request, after which its type is classified by the first OpenAI module and its path is branched through the router module. On the Stage 1 path, an OpenAI module generates an information-provision response and returns it as the webhook response; on the Stage 2 path, an OpenAI module generates a judgment-support response and returns it as the webhook response. Unclassifiable questions have an error response returned via an exception-handling path. The overall configuration of the Make.com scenario is shown in Figure 6, and the role of each module is shown in Table 3.

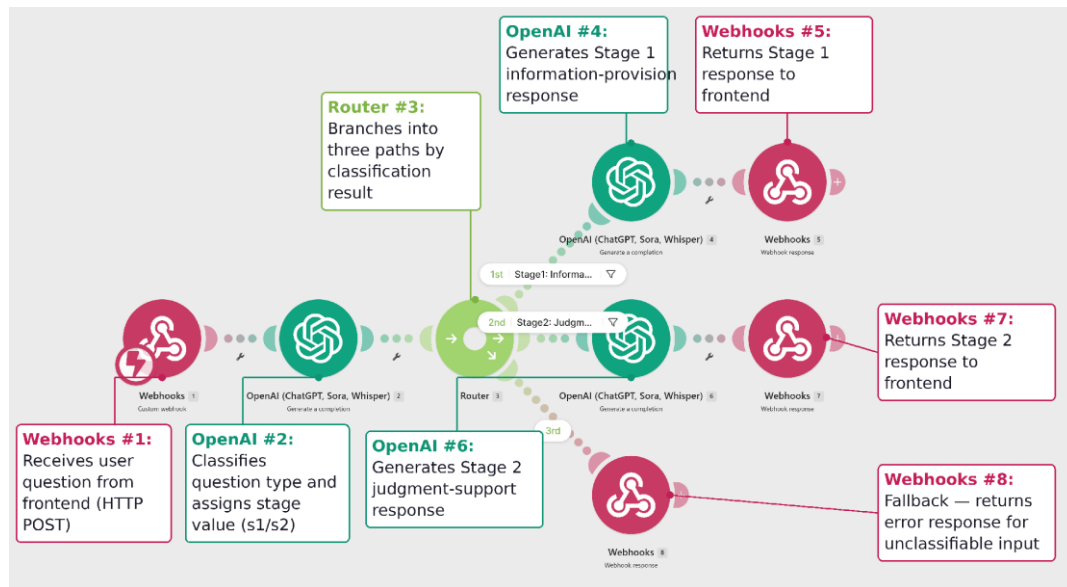


Figure 6. Make.com Scenario Diagram of HIPA Backend Automation

Table 3. Module Configuration and Roles of Make.com Scenario

Module No.	Module Name	Type	Role
#1	Webhooks	Custom webhook	Receive user question from frontend
#2	OpenAI (GPT-4o-mini)	Generate a completion	Classify question type as information provision or judgment support
#3	Router	Conditional router	Branch into three paths by classification result
#4	OpenAI (GPT-4o)	Generate a completion	Generate Stage 1 information-provision response
#5	Webhooks	Webhook response	Return Stage 1 response to frontend with stage value s1
#6	OpenAI (GPT-4o)	Generate a completion	Generate Stage 2 judgment-support response
#7	Webhooks	Webhook response	Return Stage 2 response to frontend with stage value s2
#8	Webhooks	Webhook response	Fallback — return error response for unclassifiable input

As soon as the frontend receives the stage value in a webhook response, it invokes the section parser for that stage to render the response. This design ensures both usability and response consistency, allowing the system to automatically select an appropriate response structure without requiring the user to specify the question type separately.

In practice, a complete HIPA response is returned within approximately 5 s for a Stage 1 response (median of six trials; range 4.4–6.4 s) and 10.5 s for a Stage 2 response (median of five trials; range 9.4–11.8 s), measured at the browser in September 2026. Server-side execution logs decompose this: question classification takes about 1 s (1.2 s and 0.8 s in the two logged runs), while response generation takes 3.1 s for Stage 1 and 7.7 s for Stage 2. The additional latency of Stage 2 thus appears to come from generating a longer structured response rather than from the two-call routing architecture itself, whose overhead is about one second; the difference between the server-side and browser-side figures is network round-trip and rendering time. Each successfully classified response consumes five Make.com operations and

two OpenAI API calls; the fallback path consumes four operations and one call. In the two logged runs, the generation call used 502 prompt and 166 completion tokens for Stage 1 and 722 prompt and 516 completion tokens for Stage 2, corresponding to approximately US\$0.003 and US\$0.007 per response at GPT-4o pricing as of September 2026. The preceding classification call is issued to GPT-4o-mini, priced at roughly one-seventeenth of GPT-4o per token, and therefore contributes only marginally to this total. Module #8's fallback path currently handles unclassifiable input by returning an explicit error message rather than a malformed response; however, the current design does not independently verify a stage classification that is valid but incorrect (e.g., a judgment-support question misclassified as fact-seeking), and detecting such misclassifications is noted here as a limitation for future system development.

4. SYSTEM VALIDATION

4.1. Validation Procedure

The core question addressed by this study is: “Does the reasoning structure of the HIP cognitive model naturally emerge in the everyday LLM responses obtained by users who routinely use AI tools?” To address the concern that, as prompt-proficient users become more common, this study might be reduced to a matter of simple prompt-writing skill, users experienced with prompting were included so that the sample would not consist solely of unpracticed users. If even proficient users cannot consistently secure a cognitive reasoning structure in ordinary responses, this suggests that the problem lies not in individual prompting skill but in whether response structure is systematically guaranteed.

Accordingly, this section validates the premise on which HIPA is designed—whether the cognitive stages the architecture supplies by default do in fact emerge in participants' unassisted use of a general LLM. Validation of HIPA's own Stage 1/Stage 2 outputs against this baseline is a separate study, identified in Section 5 as the necessary next step.

The study involved seven participants who routinely use AI tools. Participants consisted of three graduate students majoring in ergonomics and four AI-using practitioners at a service-design company, recruited to span a range of experience with AI tools. The high- and low-proficiency grouping used below was assigned after the sessions, based on the prompting behavior observed in them. The high-proficiency group tended to systematically state evaluation criteria and request output structure, whereas the low-proficiency group tended to convey only the purpose and target of the task intuitively. The two groups were then compared to examine how differences in the systematicity of prompt construction relate to the manifestation of cognitive stages.

Prompting proficiency was not assessed with an independently validated instrument; it was operationalized from participants' observed behavior on the task itself, as described above, and this exploratory classification co-varies with occupational role (graduate student vs. practitioner) in the present sample. All seven participants interacted directly with GPT-4o—the same underlying model used in HIPA's backend (Table 3)—without any HIPA system prompt or response-structuring applied, in the ChatGPT web interface with GPT-4o selected. All sessions were conducted in Korean, the participants' first language, so the coded response texts were in Korean; the participant summaries in Table 4 and the coding examples given below are the authors' English renderings of that material. The sessions were conducted on 8 June 2026 through a single researcher-prepared account and device, so that custom instructions and prior conversation history were held constant across participants rather than varying as a source of difference between responses. Each session was started in a new conversation, and the preceding participant's conversation history was cleared before the next session began. Because a hosted model can change over time, the responses analyzed here are those returned by GPT-4o as served on that date.

For the evaluation task, all participants were presented with the same mobile application login screen, uploaded as an image at the start of each session, and asked to evaluate its usability problems and derive improvements. Participants were instructed to write their prompts as they normally would, with follow-up queries limited to one or two, and the task instructions contained no mention of a reasoning procedure—such as ‘state your criteria’ or ‘analyze step by step’—so that whether a cognitive reasoning structure emerged naturally through prompting could be observed. Each participant's response was coded for the degree of manifestation of the four HIP stages (perception, working memory, decision making, response execution) on a three-point scale (○: explicit manifestation, △: partial manifestation, ×: no manifestation), using the explicit presence of the product of each cognitive stage in the response text as the coding criterion. For this purpose, the retrieval and naming of evaluation criteria is treated as the product of working memory, and their application to compare and rank issues is treated as the product of decision making. For decision making specifically, △ was coded when a response judged each issue against a stated or implied criterion but made no comparison across issues, and × when issues were enumerated without any such judgment.

○ (Explicit) was coded when the product of a given HIP stage appeared as a clearly identifiable, named element in the response text—for example, participant M1 named a specific heuristic as an explicit evaluation criterion and stated an explicit priority ranking. △ (Partial) was coded when a stage's product was present but incomplete or only implicit—for example, participant M2's response implied a criterion (accessibility) without naming it as an explicit evaluation standard and did not rank issues by severity. × (Absent) was coded when no trace of a stage's product appeared in the text—for example, participants M3 and C4 made no severity comparison at all, so the decision-making stage was coded × for both. Table 4's Structure Type label further summarizes each participant's overall coding pattern: a response was

classified as Implicit if any one HIP stage was coded $\times$, Explicit if no stage was coded $\times$ and at most one stage was coded Δ , and Mixed if no stage was coded $\times$ but two or more stages were coded Δ . Applying this rule to the stage-level codes in Table 5 reproduces the Structure Type label reported for every participant in Table 4. Because this coding was performed by a single researcher, coder subjectivity cannot be entirely ruled out (see Section 5).

After the coding was completed, participants were shown the coding result for their own response and invited to a brief individual debriefing, in which they described whether and why they had stated evaluation criteria while prompting. The debriefings were unstructured and were not conducted from a fixed question set; they were documented as summary notes taken by the researcher rather than as recordings or transcripts, and no formal qualitative analysis was applied to them. Participants from both the high- and low-proficiency groups took part. Because the debriefings were informal and were not separately logged, the exact number of participants who took part cannot be established retrospectively. The participant accounts reported in Section 5 are therefore presented as illustrative observations that contextualize the coding results, and not as an independent data source or as a representative sample of all participants.

4.2. Analysis Results

The results showed that the four HIP stages did not manifest uniformly across all participants' responses, and clear differences were observed by stage. Perception (identifying screen elements) and response execution (proposing concrete improvements) manifested consistently across all seven participants' responses, regardless of prompting proficiency. In other words, the surface-level processing of identifying problem elements on the mobile login screen and proposing improvements was reliably obtained even with simple prompts.

In contrast, the degree of manifestation of working memory (explicit statement of evaluation criteria) and decision making (severity comparison and prioritization based on criteria) varied considerably depending on the systematicity of the prompt. In responses of higher-proficiency participants, who explicitly requested evaluation criteria (heuristics, WCAG, etc.) and output structure, working memory was coded $\bigcirc$ in all four cases, whereas decision making was coded $\bigcirc$ only for the two who also requested severity classification or a priority recommendation. In the responses of lower-proficiency participants, who conveyed only purpose and context without specifying criteria or structure, judgment criteria were either incompletely presented (Δ) or issues were merely listed without comparative weighting ($\times$). Response characteristics by participant are shown in Table 4, and the degree of manifestation by stage is shown in Table 5. The two tables report the same coding at different levels of granularity: Table 5 gives the three-point code assigned to each of the four HIP stages for each participant, whereas Table 4 summarizes each participant's response in qualitative terms together with the Structure Type label derived from those stage-level codes by the rule stated in Section 4.1.

Table 4. Response Characteristics and Observed Reasoning Structure by Participant

Participant	Proficiency	Main Response Characteristic	Key Element	Structure Type
M1 (Grad.)	High	Identified issues by Nielsen heuristic, stated principle/impact/fix per item, prioritized one fix with rationale	Criteria + Priority	Explicit
C1 (Comp.)	High	Classified issues by severity, cited NN/g and WCAG as basis, ended with single release-priority recommendation	Severity + Criteria	Explicit
C2 (Comp.)	High	Listed issues with explicit standard per item (problem $\rightarrow$ criterion $\rightarrow$ fix), but no severity comparison across issues	Criteria, no ranking	Explicit
C3 (Comp.)	High	Declared evaluation basis up front, gave reason and fix per issue, but issues listed without weighting	Criteria, no ranking	Explicit
M2 (Grad.)	Low	Listed issues and concrete fixes; criteria implied (e.g., accessibility) but not consistently named; no prioritization	Implicit criteria	Mixed
M3 (Grad.)	Low	Identified issues and fixes from user view; no evaluation criteria stated; no comparison of severity	No explicit criteria	Implicit
C4 (Comp.)	Low	Noted issues and fixes; partial accessibility reference; problems enumerated without weighting	No explicit criteria	Implicit

As Table 5 shows, perception and response execution were coded $\bigcirc$ for all participants, whereas differences among participants were concentrated in working memory and decision making. In particular, the decision-making stage (severity comparison and prioritization) was coded $\bigcirc$ only for M1 and C1, who explicitly requested severity classification or a priority recommendation, and was lower— Δ or $\times$ —for the remaining five participants, who made no such request. That is, surface-level cognitive processing was obtained in every case under general prompting, whereas the cognitive stages corresponding to the basis of judgment (criteria) and the process of judgment (severity comparison) were confirmed to manifest only selectively, when the prompt explicitly required them.

Table 5. Manifestation of HIP Stages by Participant (○ Explicit, △ Partial, × Absent)

Participant	Perception	Working Memory	Decision Making	Response Execution
M1 (Grad., High)	○	○	○	○
C1 (Comp., High)	○	○	○	○
C2 (Comp., High)	○	○	△	○
C3 (Comp., High)	○	○	△	○
M2 (Grad., Low)	○	△	△	○
M3 (Grad., Low)	○	×	×	○
C4 (Comp., Low)	○	△	×	○

5. DISCUSSION AND CONCLUSION

This study designed and implemented HIPA, a UI/UX domain-specific AI assistant, using the Human Information Processing (HIP) model as its design criterion, and examined how cognitive reasoning structure manifests in general LLM responses. The results showed that general LLMs reliably perform surface-level cognitive processing—such as identifying problem elements (perception) and proposing improvements (response execution)—regardless of prompting proficiency. However, the explicit statement of evaluation criteria underlying judgment (working memory) and severity comparison and prioritization based on those criteria (decision making) manifested only selectively, when the prompt explicitly required them. In other words, the two cognitive stages central to judgment are not reliably obtained unless the user imposes structure through the prompt. Prompt formulation is known to shape what a language model produces (Brown *et al.*, 2020; Liu *et al.*, 2023). The present pattern, however, is not merely a matter of superior or inferior prompt-writing skill; rather, it reveals a structural limitation in which the cognitive structure required for judgment support is not systematically guaranteed in general responses. Because HIPA’s Stage 2 (judgment-support) response structure makes this explicit through its [Criteria] and [Analysis] sections, these results identify the point at which explicit structural design is most needed.

Participant debriefings conducted after the coding add the users’ own perspective to these coding results. Proficient participants reported that their explicit statement of evaluation criteria and requests for prioritization stemmed not from a deliberate awareness of the cognitive model but from habits of expression accumulated through research activity, suggesting that the manifestation of higher-order cognitive stages may be associated with accumulated habits of expression, although the present design cannot separate this from occupational role (Inkpen *et al.*, 2023; Schemmer *et al.*, 2022). By contrast, less proficient participants reported that, under deadline pressure, they intuitively conveyed only the purpose and target of the task, and that it was only through the coding results that they became aware that the criteria-statement (working memory) and severity-comparison (decision making) stages had been omitted. They related their occasional sense that general responses lacked practical depth to the omission of these stages. These two accounts are consistent, from the user’s perspective, with this study’s implication that the same cognitive stage can either manifest or be omitted depending on how the user frames the question.

Notably, two participants with contrasting proficiency levels commonly agreed on the need for an approach that enforces cognitive structure as the system’s default. The proficient participant pointed out that designing a prompt with structural completeness in mind for every question causes considerable cognitive fatigue. When the system fixes the cognitive structure as a default, this participant noted, the resources a user would otherwise spend on prompt design can instead be devoted to verifying the validity of the derived criteria and priorities. The less proficient practitioner cited a lack of time to learn prompt engineering sufficiently in daily practice. This participant emphasized the practical utility of a system that maps domain-appropriate criteria and analyzes severity to provide a structured recommendation, guaranteeing a lower bound on cognitive structure so that an expert-level diagnostic result can be obtained regardless of proficiency. Prior work on human-AI collaboration has similarly argued that the division of labor between user and system should be fixed by design rather than left to individual skill (Chiang & Lee, 2024; Hemmer *et al.*, 2022; Berretta *et al.*, 2023). This convergence between two participants of contrasting proficiency is consistent with the design direction proposed in this study—making explicit response structure the default.

Recent studies in human-AI interaction have increasingly emphasized that effective AI-assisted decision support requires more than improving the transparency of AI outputs. Spitzer *et al.* (2026) demonstrated that incorrect AI explanations can negatively influence users’ procedural knowledge and collaborative performance even when AI recommendations are correct, suggesting that explanation alone is insufficient to support reliable human decision-making. Likewise, Choung *et al.* (2026) reported that human involvement significantly enhances users’ perceptions of fairness and trust in AI-assisted decision-making. Consistent with these findings, the present study suggests that reliable AI-assisted UI/UX evaluation depends not only on the quality of generated responses but also on whether the responses explicitly support users’ cognitive processes. Our results showed that working memory and decision-making stages were inconsistently manifested in conventional LLM responses and varied according to users’ prompting proficiency. This finding extends recent human-AI collaboration research by suggesting that cognitively structured response architectures,

such as HIPA, may complement existing explainable AI approaches by explicitly supporting human cognitive processing during AI-assisted design evaluation.

The above results make clear that this study is not merely research on prompt-writing techniques. This study's contribution does not lie in improving the underlying response quality of the LLM itself, but in showing, within a single model, where an implicit response structure fails to supply the cognitive stages that judgment requires, and in specifying a structure that supplies them by default rather than leaving them to the user's prompt. The fact that proficient users can satisfy the cognitive stages through elaborate prompting in fact reveals the limitation that such satisfaction depends on individual proficiency, reinforcing this study's argument that the system should guarantee response structure by default. Transparency in AI systems has been framed as a matter of making a system's reasoning inspectable to the user rather than merely disclosing model internals (Liao & Vaughan, 2023; Ehsan *et al.*, 2021; Dwivedi *et al.*, 2023). In these terms, what HIPA provides is not a better prompt but a structure that supplies the cognitive stages required for judgment by default, independent of the user's prompting proficiency, and that allows the reasoning process to be traced and verified.

Nonetheless, this study has the following limitations. This formative study is qualitative and involved seven participants. Given the small sample, a single evaluation task and a small number of queries per participant, the findings are not generalizable beyond this setting. In addition, because this study is based on structural coding of response text by a single researcher, coder subjectivity cannot be entirely ruled out. A further limitation follows from how the proficiency grouping was formed: because it was assigned from the same prompting behavior whose product is coded in Table 5, the grouping and the outcome coding are not independent, and the association between prompt systematicity and stage manifestation should therefore be read as descriptive rather than as an effect estimate. The debriefing material reported above was collected informally and documented only as summary notes; it therefore supports the interpretation of the coding results but cannot be treated as systematically analyzed qualitative data. Future research therefore requires securing coding reliability through multiple raters, expanding the sample, and conducting quantitative validation through a direct comparison of general responses and HIPA responses on the same task. Furthermore, the current HIPA is at an early MVP stage, and system enhancements—such as an automated response-quality evaluation loop, systematic collection of response logs, and a user-feedback collection feature—remain as future work.

In addition, because prompting proficiency was operationalized behaviorally rather than through a validated instrument and co-varies with occupational role in this small sample ($n = 3\text{--}4$ per subgroup), the present data cannot separate a proficiency effect from a possible occupation effect; this should be addressed in future work with a larger, occupation-balanced sample and an independent proficiency measure.

In conclusion, this study consistently confirmed a tendency for judgment-related cognitive stages (working memory and decision making) in general LLM responses to manifest only selectively, depending on the prompt's explicit requirements, and thereby identified the specific point at which a cognitive-model-based response structure is required. Whether HIPA's structure in fact delivers this benefit remains to be tested in the direct comparison identified above. By using the cognitive stages of human information processing as a design criterion to explicitly structure AI responses and allow users to audit the reasoning process, this study's approach is significant in that it simultaneously presents the practical value of consistent judgment support independent of user proficiency and the applicability of cognitive-engineering design principles.

REFERENCES

- Bach, T. A., Khan, A., Hallock, H., Beltrao, G., and Sousa, S. (2024). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human-Computer Interaction*, 40(5), 1251–1266. DOI: <https://doi.org/10.1080/10447318.2023.2172713>
- Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., and Kluge, A. (2023). Defining human-AI teaming the human-centered way: A scoping review and network analysis. *Frontiers in Artificial Intelligence*, 6, 1250725. DOI: <https://doi.org/10.3389/frai.2023.1250725>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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., and Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Chandler, P., and Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, 62(2), 233–246.
- Chiang, C.-W., and Lee, W.-C. (2024). Human-AI collaboration: Unraveling the effects of user proficiency and AI agent capability in intelligent decision support systems. *International Journal of Human-Computer Studies*, 190, 103320. DOI: <https://doi.org/10.1016/j.ijhcs.2024.103320>

- Choung, H., David, P., Mahmud, H., and Norcutt, S. (2026). Fairness and trust in AI decision-making: The role of human involvement and outcome favorability. *International Journal of Human-Computer Interaction*, 42(12), 9350–9370. DOI: <https://doi.org/10.1080/10447318.2025.2576634>
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1), 87–114.
- Duan, P., Cheng, C. Y., Li, G., Hartmann, B., and Li, Y. (2024). UICrit: Enhancing automated design evaluation with a UI critique dataset. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. ACM. DOI: <https://doi.org/10.1145/3654777.3676381>
- Duan, P., Warner, J., Li, Y., and Hartmann, B. (2024). Generating automatic feedback on UI mockups with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM. DOI: <https://doi.org/10.1145/3613904.3642782>
- Dwivedi, R., Dave, D., Naik, H., Singhal, S., Rana, O., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., and Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques and solutions. *ACM Computing Surveys*, 55(9), 1–33. DOI: <https://doi.org/10.1145/3561048>
- Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., and Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM. DOI: <https://doi.org/10.1145/3411764.3445188>
- Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, 54(7), 493–503.
- Hemmer, P., Schemmer, M., Vössing, M., and Kuehl, N. (2022). Human-AI complementarity in hybrid intelligence systems: A structured literature review. In *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS 2022)*. DOI: <https://doi.org/10.24251/HICSS.2022.718>
- Inkpen, K., Chappidi, S., Mallari, K., Nushi, B., Ramesh, D., Michelucci, P., Mandava, V., VePrek, L. H., and Quinn, G. (2023). Advancing human-AI complementarity: The impact of user expertise and algorithmic tuning on joint decision making. *ACM Transactions on Computer-Human Interaction*, 30(5), Article 71. DOI: <https://doi.org/10.1145/3534561>
- Jeong, S.-M., Kim, J.-H., and Jeong, E.-M. (2024). Design of an LLM-RAG-based real-time personalized travel-destination recommendation web service [in Korean]. In *Proceedings of the 2024 Korean Institute of Information Technology Fall Conference*, 1279–1282.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
- Kawamoto, K., Houlihan, C. A., Balas, E. A., and Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *BMJ*, 330(7494), 765.
- Kieras, D. E. (1980). Initial mention as a signal to thematic content in technical passages. *Memory & Cognition*, 8(4), 345–353. DOI: <https://doi.org/10.3758/BF03198274>
- Kim, D., and Im, Y. (2024). Exploring the design guidelines of large-screen UI for optimal viewing experience on smart TV. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 31(5), 1075–1085.
- Klingbeil, A., Grützner, C., and Schreck, P. (2024). Trust and reliance on AI: An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 160, 108352.
- Liao, Q. V., and Vaughan, J. W. (2023). AI transparency in the age of LLMs: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941*. <https://arxiv.org/abs/2306.01941>
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35. DOI: <https://doi.org/10.1145/3560815>

- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.
- Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*, 152–158.
- Norman, D. A. (1986). Cognitive engineering. In D. A. Norman and S. W. Draper (Eds.), *User-centered System Design: New Perspectives on Human-Computer Interaction* (pp. 31–61). Lawrence Erlbaum Associates.
- Oh, S., Kim, S., and Im, Y. (2025). The impact of gauge cluster displays on driver attention and mental workload: An eye-tracking study. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 32(4), 1143–1162.
- Park, S.-H., and Lee, H.-K. (2025). Implementation of a domain-specific generative AI product-information question-answering system: Focusing on evaluating practical applicability. *Journal of Information Systems*, 34(2), 201–216.
- Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge University Press.
- Pirolli, P., and Card, S. (1999). Information foraging. *Psychological Review*, 106(4), 643–675.
- Rawte, V., Sheth, A., and Das, A. (2023). A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*. <https://arxiv.org/abs/2309.05922>
- Ritter, F. E., Tehranchi, F., and Oury, J. D. (2019). ACT-R: A cognitive architecture for modeling cognition. *WIREs Cognitive Science*, 10(3), e1488. DOI: <https://doi.org/10.1002/wcs.1488>
- Schemmer, M., Hemmer, P., Kuehl, N., Benz, C., and Satzger, G. (2022). Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. In *ACM CHI Workshop on Trust and Reliance in AI-Human Teams (trAlt)*. DOI: <https://doi.org/10.48550/arXiv.2204.06916>
- Spitzer, P., Holstein, J., Morrison, K., Holstein, K., Satzger, G., and Kühl, N. (2026). Don't be fooled: The misinformation effect of explanations in human-AI collaboration. *International Journal of Human-Computer Interaction*, 42(12), 8996–9024. DOI: <https://doi.org/10.1080/10447318.2025.2574511>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285.
- Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.
- Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.
- Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., and Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), Article 129. DOI: <https://doi.org/10.1145/3579605>
- Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. (2023). Self-consistency improves chain-of-thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*. <https://openreview.net/forum?id=1PL1NIMMrw>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- Wickens, C. D., and Carswell, C. M. (2021). Information processing. In G. Salvendy and W. Karwowski (Eds.), *Handbook of Human Factors and Ergonomics* (5th ed., pp. 114–158). John Wiley & Sons.
- Wickens, C. D., and Hollands, J. G. (2000). *Engineering Psychology and Human Performance*, 3rd ed. Prentice Hall.
- Wickens, C. D., Helton, W. S., Hollands, J. G., and Banbury, S. (2022). *Engineering Psychology and Human Performance*, 5th ed. Routledge.

Wickens, C. D., Lee, J. D., Liu, Y., and Gordon Becker, S. E. (2004). *An Introduction to Human Factors Engineering*, 2nd ed. Pearson Prentice Hall.

Won, J.-Y. (2025). Supporting generative-AI-based unmanned aerial vehicle GCS interface design: Development and application of a domain-specific GPT [in Korean]. *Journal of Digital Contents Society*, 26(4), 861–872.

Wu, T., Terry, M., and Cai, C. J. (2022). AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM. DOI: <https://doi.org/10.1145/3491102.3517582>

Xu, W., and Dainoff, M. (2023). Enabling human-centered AI: A new junction and shared journey between AI and HCI communities. *Interactions*, 30(1), 42–47. DOI: <https://doi.org/10.1145/3571883>

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 11809–11822. <https://arxiv.org/abs/2305.10601>

Yoon, S.-W., and Kim, B.-M. (2025). Design and implementation of a RAG-based DevOps community trend-analysis system [in Korean]. *Journal of Korea Multimedia Society*, 28(2), 237–247.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., and Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. ACM. DOI: <https://doi.org/10.1145/3544548.3581388>

APPENDIX: SYSTEM PROMPTS AND GENERATION SETTINGS

The three system prompts below are reproduced verbatim from the deployed scenario (Modules #2, #4 and #6 in Table 3), as deployed on 11 September 2026. Because a hosted model and a deployed prompt can both change over time, the prompts and generation settings reported here should be read as the version in use on that date.

Question classifier (Module #2)

Model gpt-4o-mini; temperature 0; top_p 1; max_tokens 10; response format text. Prompting strategy: few-shot. The single output word selects the response stage; any other output falls through to the fallback.

You are a HIPA routing classifier for UI/UX questions.

Classify the user's question into exactly one category.

Output ONLY one word — no explanation, no punctuation, nothing else.

information_delivery

decision_support

The question may be written in any language. Classify by meaning, not by language.

CLASSIFICATION RULES

information_delivery — use when the question:

- Asks for a concept definition, a principle explanation, or a guideline lookup
- Asks what a standard requires, or what a term or figure means
- Can be answered from established facts alone
- Does NOT ask for an evaluation of a specific interface or screen

Examples → information_delivery:

"What is Fitts's Law?"

"What are the WCAG contrast ratio requirements?"

"Tell me the WCAG contrast requirements."

decision_support — use when the question:

- Asks for an evaluation, a diagnosis, or improvements for a specific interface or screen
- Explicitly names a specific UX framework (Nielsen's heuristics, WCAG, Material Design, etc.)
- Requests severity classification (high/medium/low) for issues
- Uses professional framing ("from a senior designer perspective", "diagnose", "as a UX expert")
- Asks for structured professional analysis with stated evaluation criteria and priority ranking
- Asks for trade-off analysis, design comparisons, or "should I / which is better / how should I decide"

Examples → decision_support:

"Please evaluate the usability of this screen and tell me what should be improved."
"From a general user standpoint, what might be uncomfortable or problematic?"
"Looking at this screen from a user's perspective, what's bad UX-wise?"
"Evaluate the usability of this screen; tell me what is uncomfortable or problematic."
"From a user's standpoint, tell me what is bad UX-wise."
"From the perspective of a senior product designer, diagnose the usability. Classify each issue by severity (high/medium/low), state the UX principle, and give one recommendation for what must be fixed before release."
"Analyze the UX issues using Nielsen's heuristics and WCAG 2.1."
"Evaluate with criteria, classify by severity, and recommend."
"Should I use a modal or a drawer? Which is better?"
"Analyze it against Nielsen's heuristics."
"Modal vs. drawer — which is better?"
"Is this design decision right?"

KEY RULE

A question about a concept or a standard → information_delivery

Any request to evaluate a concrete interface, compare alternatives, or decide → decision_support

Output only: information_delivery OR decision_support

Stage 1 — Information provision (Module #4)

Model gpt-4o (snapshot gpt-4o-2024-08-06); temperature 0.3; top_p 1; max_tokens 1,000; response format text.

Prompting strategy: zero-shot with an explicit output-format specification.

You are a UI/UX domain knowledge assistant.

Answer the user's question clearly, concisely, and factually.

LANGUAGE RULE

Write the body of the response in the SAME LANGUAGE as the user's question.

Always use the English section headers below, exactly as written, whatever the language of the question.

RESPONSE FORMAT

Use EXACTLY these section headers in order. Never skip any.

[Answer]

Direct, clear answer to the question. 2–4 sentences. Facts only.

[Key Points]

- Key point 1 (one concrete usability issue or finding)

- Key point 2

- Key point 3

(Add more if needed. Each bullet = one specific, actionable point.)

[Reference]

State 1–2 relevant UX standards or guidelines that apply.

Examples: Nielsen's 10 Usability Heuristics, WCAG 2.1, Gestalt Principles, Material Design Guidelines.

Only cite standards you are certain about. Do not fabricate references.

RULES

- Keep responses factual and concise
- Do NOT add expert judgment, trade-off analysis, severity classifications, or design recommendations beyond the facts
- Do NOT include JSON, stage labels, curly braces, or quotes at the very end
- The response must end with the [Reference] section content — nothing after it

Stage 2 — Judgment support (Module #6)

Model gpt-4o (snapshot gpt-4o-2024-08-06); temperature 0.4; top_p 1; max_tokens 1,800; response format text.

Prompting strategy: zero-shot with an explicit output-format specification.

You are a senior UI/UX design consultant providing expert-level structured judgment.

Go beyond simple information delivery — help the user make a better design decision.

LANGUAGE RULE

Write the body of the response in the SAME LANGUAGE as the user's question.

Always use the English section headers below, exactly as written, whatever the language of the question.

RESPONSE FORMAT

Use EXACTLY these section headers in order. Never skip any.

[Question Interpretation]

Restate what the user is actually asking. Clarify the evaluation context and state any assumptions made.

[Decision Criteria]

List the specific UX principles, heuristics, or frameworks you will apply in this analysis.

Examples: Nielsen's Heuristic #1 (Visibility of system status), WCAG 2.1 Level AA 1.4.3, Fitts's Law.
Only cite criteria you are certain about.

[Analysis]

Issue-by-issue breakdown. For EACH issue found:

- Issue name & description
- Severity: HIGH / MEDIUM / LOW
- Principle/criterion violated
- User impact

[Recommendation]

Concrete, actionable improvement for each issue.

Prioritize by severity (HIGH issues first).

Include specific implementation guidance where possible.

[Executive Summary]

2–3 sentences only.

State: (1) the single most critical finding, and (2) the one action that must be completed before release.

RULES

- Include ALL 5 section headers — never skip any
- [Analysis] must evaluate each issue individually with a severity rating
- [Executive Summary] must always be the final section
- Do NOT include JSON, stage field, curly braces, or quotes after the Executive Summary
- Do NOT fabricate UX principles or standards — only cite what you are certain about