ACM GreekCHI-CYCHI Symposium 2026

Supporting UI designer's decisions with AI RAG assistance

Marina Maciá | marinamf@dlsi.ua.es - **David Rizo** | drizo@dlsi.ua.es
University of Alicante - EASDA - ISEA.CV

1. Introduction

A good practice in interface design is to base decisions on relevant standards, style guides, and reference materials. However, these collections of documents are often so extensive that reviewing them manually can be very time-consuming. To speed up this process, it is common to ask for advice from commercial large language models (LLMs) with conversational capabilities.

1.1 Problem

- These LLMs generate a response regardless of their actual knowledge, and their references become post-hoc attributions.
- It is often not possible to attach the documents, because licensing terms and internal policies prohibit their use in third-party services.
- Some designers are cautious about sharing their work, since there's no way to verify where the uploaded content ends up.

1.2 Research objective

RQ *What would change if the assistant ran on the designer's computer and responded from a reference library chosen by them, citing each statement?*

- Support interface design decisions with useful and verifiable responses.
- Let the designer verify each statement by opening the cited page, and keep exploring the source.
- Keep licensed material on the designer's machine, with no data sent to external services.

2. State of the art

- UI research community has focused on various phases of design process: creativity support tools, prototype, generating code, among others.
- Several plugins for design tools highlight interest in supporting this phase such as checking WCAG 2.2 rules or compliance with a design system, but they cannot answer the open questions that arise while the decision is still being made.
- Retrieval has already been applied in domains where verifying the source is part of the task, such as in medicine, law or art context.

3. Methodology

This study compares four approaches to solving the same design query under the real-world conditions of a small-scale study.

3.1 Consultation stack

Step 1 **Corpus**

184 native digital and scanned documents (ISO standards on usability and accessibility, interface design guidelines, reference literature from various design disciplines).

Step 2 **Ingestion**

The content of each page is retrieved from two ways (Fig. 1).

- **Textual**: The text of each page is extracted using a cascade of methods with increasing cost. The resulting text is split into chunks, encoded, and indexed in a text vector database. Figures are described in natural language and are also indexed.
- **Visual**: Each page is directly encoded as an image using a late-interaction multivector vision-language retrieval system and indexed in a visual vector database.

Step 3 **Model selection**

The entire system runs on a single consumer-grade GPU (24 GB): Gemma 4 generates the answers, nomic-embed retrieves text and ColQwen2.5 retrieves page images. Fixed parameters make the experiment deterministic.

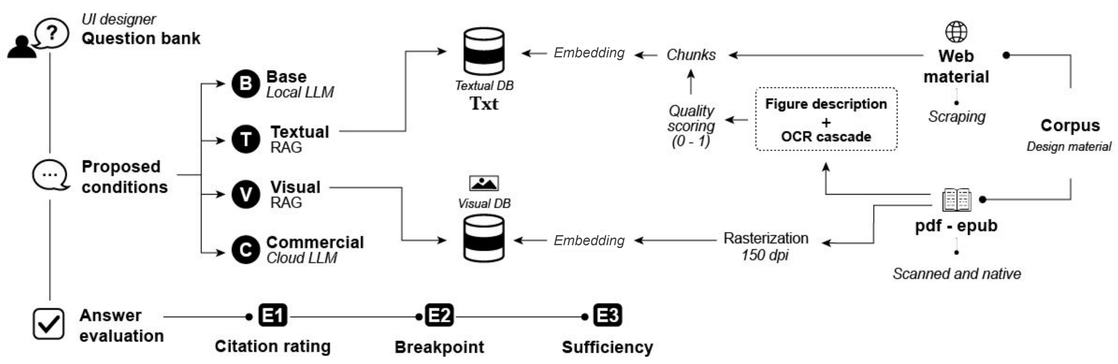


Figure 1: Proposal overview

3.2 Evaluation protocol

To answer the research question, this study compares four approaches:

- **B) Base model**: The local model answers from its own knowledge.
- **T) Textual RAG**: The model receives the ten most relevant text fragments from the text vector DB.
- **V) Visual RAG:** The ten most relevant pages from the visual vector DB are attached as images.
- **C) Commercial system**: Anthropic Claude Fable without access to the corpus, as a commercial reference.

Question bank: An expert designer wrote 48 questions from her own professional practice and indicated which pages in the corpus answer them.

Blind evaluation: The four responses to each question are shown in random order and without labels, and each cited page can be opened (Fig. 2). The comparison is paired, so each question serves as its own control.

Two ratings per response: Sufficiency (on a graded scale) and citation accuracy (whether each cited source supports the claim).

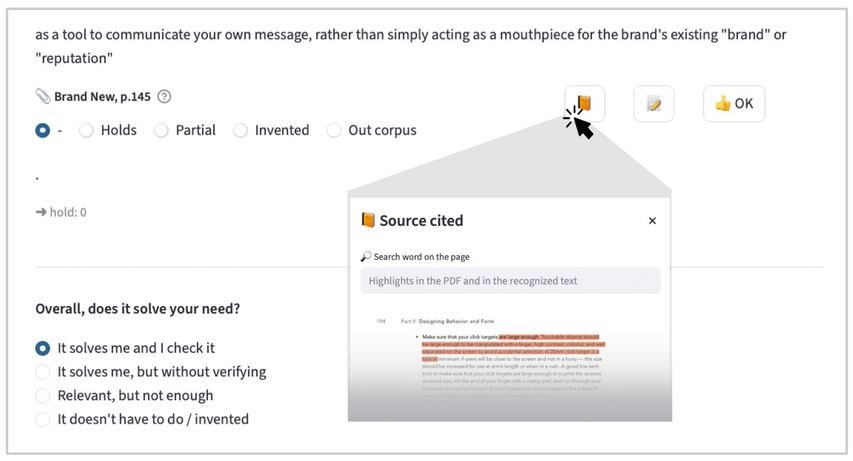


Figure 2: Web-based interface for evaluation

4. Results and discussion

- Retrieval did not make the local model resolve more questions, but it made its answers verifiable. B, T and C resolve almost the same number of questions, but only in T could every answer be verified.
- In T, 0.94 of the citations support the claim they are attached to.
- V finds the source but does not use it. It locates the expected page as often as T, but its answers fail to make use of it.
- On questions without a single correct answer, B and C give a categorical but unverifiable verdict, while T and V surface authors who disagree. The designer can then decide on evidence rather than on trust.

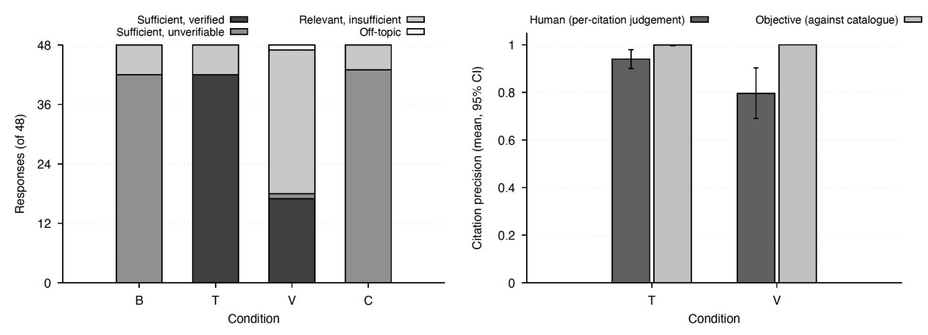


Figure 3: (a) Sufficiency by condition, in number of responses out of 48. (b) Human and objective citation precision for T and V, averaged per response. Error bars are 95% confidence intervals across responses.

5. Conclusions and future work

Following the problem-solving view of HCI research, we interpret these results as new capabilities: What a designer can do now that she could not do before. Our central contribution is making LLM answers verifiable, by grounding them in a personal corpus that the model draws on to answer questions.

Future Work: Since these results are based on the opinion of a single expert, future work will involve incorporating multiple evaluators, integrating the study into ongoing design tasks, and improving the visual component.

References

Duan, P., Warner, J., Li, Y., & Hartmann, B. (2024). Generating automatic feedback on UI mockups with large language models. Proc. CHI.

Faysse, M., et al. (2025). ColPali: Efficient document retrieval with vision language models. Proc. ICLR.

Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. Proc. EMNLP.

Ji, Z., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12)

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS.

Liu, N. F., Zhang, T., & Liang, P. (2023). Evaluating verifiability in generative search engines. Findings of EMNLP.

Oulasvirta, A., & Hornbæk, K. (2016). HCI research as problem-solving. Proc. CHI.

This work is supported by the Spanish Ministry of Science, Innovation and Universities through the LEMUR project (PID2023-148259NB-I00) and ISEACV SIGMA project.