Journal of Artificial Intelligence and Computing Applications (2025) - **Special Issue** - 3(2): 38–43

MAIKRON

Journal of Artificial Intelligence and Computing Applications

Expanded abstract

On-device conversational agent for psycho-oncology based on acceptance and commitment therapy manuals

Samara Acosta-Jiménez[1], **Miguel M. Mendoza-Mendoza**[1], **Gerardo N. Rivera-Rojas**[1], **Jorge I. Galván-Tejada**[1], **José M. Celaya-Padilla**[1], **and Carlos E. Galván-Tejada**[1,*]

[1]Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Mexico

ABSTRACT

Psychological distress is highly prevalent among individuals diagnosed with cancer, yet access to specialist psycho-oncology services remains limited by workforce shortages, geographical barriers, and concerns about privacy. This study presents a fully *on-device* conversational agent that combines Retrieval-Augmented Generation with a curated corpus of Acceptance and Commitment Therapy (ACT) manuals to deliver evidence-grounded emotional support without reliance on cloud resources. Multiple ACT guides and workbooks are parsed using *PyMuPDF*, segmented into 250-character chunks via a token-aware recursive splitting strategy, and embedded with *all-MiniLM-L6-v2* sentence transformers. The resulting are indexed in a FAISS *IndexFlatIP* store. At inference, LangChain retrieves and ranks the top-k passages, while LangGraph enforces source fidelity before passing the context to a locally hosted `llama3` model served via `ChatOllama`. Preliminary interactions suggest that the agent delivers concise, empathic responses referencing core ACT processes, while attaching inline citations that trace each claim to a specific passage. All responses are generated in real time on CPU-only hardware, preserving user privacy and making the system viable in low-resource clinical environments. Although current evaluation remains qualitative, no hallucinated citations or clinically unsafe statements are observed, indicating robust factual grounding at this early stage. Future work will add psycho-oncology texts, enable optional web search for unseen queries, and run concordance studies with psycho-oncologists to measure accuracy, tone, and usability. This lightweight offline pipeline therefore paves the way for privacy-preserving chatbots that enhance psychosocial care in oncology and other mental-health settings.

Keywords: cancer, chatbot, mental health

1. Introduction

Cancer constitutes a pressing global health challenge: the World Health Organization reports 20 million new diagnoses and 9.7 million deaths in 2022, while five-year prevalence already exceeds 53 million survivors and annual incidence is projected to rise above 35 million by 2050 [1, 2]. Beyond its somatic burden, the disease inflicts a profound psychological toll. Meta-analytic evidence indicates that roughly one third of adults with cancer meet criteria for an anxiety, depressive, or adjustment disorder [3], and narrative syntheses confirm that up to one half of patients with advanced disease experience clinically significant distress [4]. Such symp-

E-mail address: ericgalvan@uaz.edu.mx
https://doi.org/10.5281/zenodo.17237143
 This article is part of the Special Issue on ICAIMH 2025. ISSN: 3061-8843

toms translate into poorer treatment adherence, increased unplanned hospitalizations, diminished quality of life, and an 85% elevation in suicide mortality relative to the general population [5]. Although international guidelines mandate routine psychosocial screening, workforce shortages, geographical barriers, financial constraints, and stigma restrict access; fewer than one-fifth of distressed patients ultimately receive specialized mental-health care [6, 7, 8].

Digital interventions provide a scalable means to narrow this gap. Web- and app-based programs yield modest yet significant reductions in depression and anxiety, but engagement often wanes once human guidance is withdrawn. Fully automated conversational agents offer real-time empathic interaction; however, generic large-language-model chatbots are vulnerable to hallucination and lack domain grounding—limitations that pose unacceptable risks in oncology. Retrieval-Augmented Generation (RAG) mitigates these concerns by coupling neural text generation with dense semantic retrieval so that every output is explicitly supported by external evidence [9]. Early mental-health chatbots such as Woebot demonstrate that cognitive-behavioral strategies can be delivered safely and effectively through text-based conversation [10], yet no system to date targets the complex psychosocial needs of people living with cancer using a rigorously grounded RAG architecture.

Acceptance and Commitment Therapy (ACT)—a third-wave behavioral intervention centered on psychological flexibility—offers a mechanism-based framework ideally suited to this purpose. ACT integrates six interlocking processes (acceptance, cognitive defusion, present-moment awareness, self-as-context, values clarification, and committed action) and has accumulated robust empirical support across chronic-illness populations, including oncology.

This study introduces an innovative on-device chatbot that couples RAG with a multilingual corpus of ACT manuals. By fusing rapid semantic retrieval with contextualized natural language generation, the system delivers empathic, evidence-backed answers without cloud reliance, reducing privacy and hallucination risks. The tool positions itself as a scalable adjunct for psycho-oncology support; its effectiveness and safety will be assessed through rigorous clinical validation in future work.

2. Methodology

The notebook implements a RAG workflow executed entirely on-device. Figure 1 shows the LangGraph control flow: text chunks are embedded, stored in a local `FAISS` index, retrieved on demand, and passed to a locally served `llama3` model via `ChatOllama`.

To embed this evidence-based approach into an accessible digital tool, the present study curates a multilingual ACT corpus extracted from clinician-endorsed guides and manuals, including *Terapia de Aceptación y Compromiso* [11], the therapist and patient workbooks *Duelo: Tratamiento basado en ACT* [12], and Hayes' *Proceso del Cambio Consciente* [13]. To transform the curated ACT corpus into a machine-readable knowledge base, each PDF is parsed with `PyMuPDF`, cleaned, and split into 250-character chunks via a `RecursiveCharacterTextSplitter`. Every chunk is embedded as a 384-dimensional sentence vector using `all-MiniLM-L6-v2`. The resulting matrix is stored in a `FAISS IndexFlatIP` structure—FAISS (Facebook AI Similarity Search) is an open-source C++/Python library from Meta that performs sub-millisecond nearest-neighbour search in dense-vector spaces even on commodity CPUs, making it ideal for strictly on-device deployments[14].

When a question arrives, the system collects the most relevant text snippets and removes any that fail a basic quality check. The remaining snippets are sent to `llama3`, Meta's freely available language model[15], running locally through the lightweight `ChatOllama` program. This setup is chosen because it (i) keeps all data on the device for maximum privacy, (ii) avoids ongoing cloud costs, and (iii) connects directly to the retrieval code through LangChain's built-in ChatOllama link [16].

The methodology is presented as a fully on-device RAG pipeline. First, the ACT manuals undergo cleansing, segmentation into 250-character chunks, embedding with `all-MiniLM-L6-v2`, and storage in a `FAISS` index. Next, a `LangChain/LangGraph` workflow retrieves the most relevant passages, verifies their relevance, and forwards them to a locally hosted `llama3` model served through `ChatOllama`, which generates citation-grounded responses. Finally, the approach outlines the hallucination-mitigation guardrails, an optional web-search extension slated for future work, and the precise steps required to reproduce the experiment offline.

2.1 Imported libraries

The code relies on a deliberately small dependency set, so the prototype can be replicated on any machine with `pip`, Python's built-in package installer that fetches and installs libraries from the Python Package Index (PyPI).

Table 1 lists every third-party package referenced in an `import` statement; any module not shown belongs to the Python standard library. All packages install cleanly on Python 3.10 with no GPU support required.

2.2 Corpus preparation

Seven ACT manuals and workbooks (PDF) form the knowledge base. Each file is opened using `fitz`; headers, footers, and non-text objects are removed, yielding clean UTF-8 text. A `RecursiveCharacterTextSplitter` (`chunk_size=250`, `chunk_overlap=0`) divides the text into fixed-length segments that preserve local coherence while remaining within the context window of the target model. Every segment is embedded as a 384-dimensional vector using `HuggingFaceEmbeddings("all-MiniLM-L6-v2")`

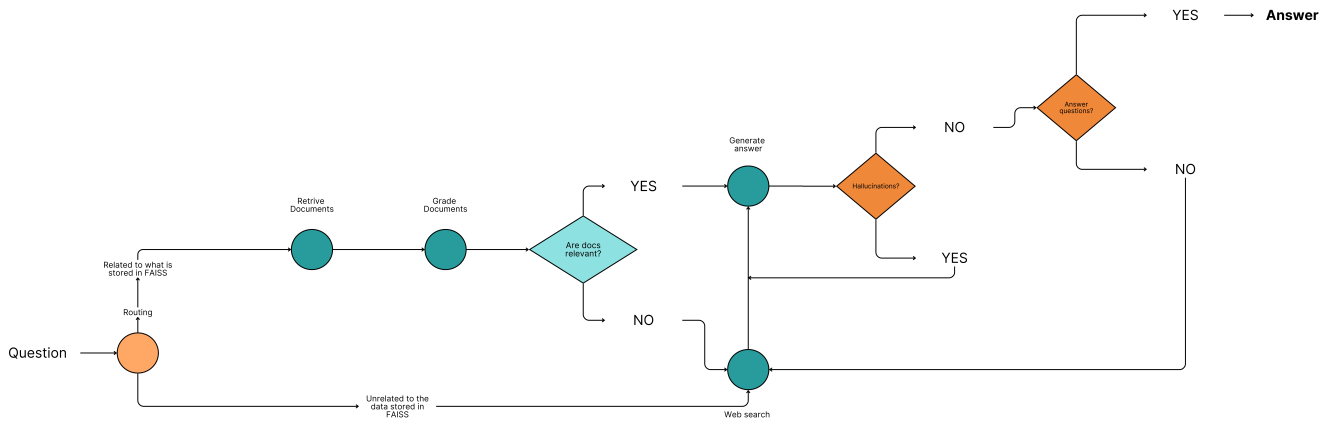

Figure 1. LangGraph control flow used in the notebook. "Retriever" queries FAISS; "LLM" invokes ChatOllama.

Table 1. External Python packages referenced in the notebook

Package	Role
PyMuPDF (*fitz*)	PDF text extraction
python-dotenv	Load `.env` variables
TextBlob	Lightweight text inspection
sentence-transformers	Embeddings (`all-MiniLM-L6-v2`)
faiss-cpu	Vector index (`IndexFlatIP`)
langchain	Core RAG primitives
`langchain_community`	FAISS wrapper, ChatOllama client
`langchain_core`	Prompt/schema abstractions
`langchain_huggingface`	HF embedding adapter
langgraph	Branching and validation nodes
ollama	Serve local `llama3` model
tavily_search	Optional web-search tool

and inserted into a `faiss.IndexFlatIP` index, enabling cosine-similarity retrieval directly on CPU with no cloud dependencies.

2.3 Retrieval and answer generation

1. *Router* – a keyword filter routes oncology and mental-health questions to the ACT corpus; unrelated queries return an *out-of-scope* message.
2. *Retriever* – `vectorstore.as_retriever()` returns the top-k = 4 passages; no similarity threshold is applied.
3. *Document grader* – a structured prompt plus `ChatOllama(model="llama3", temperature=0, format="json")` assigns a relevance score (1–5) to each passage; passages rated < 4 are discarded.
4. *Answer generator* – the filtered passages are passed to the same model to produce a plain-language answer that includes inline citations.

2.4 Hallucination guard

A validation node re-queries the `FAISS` index; if any citation string in the generated answer does not match the retrieved passages, the response is discarded and regenerated using a narrower context window.

2.5 Reproducibility

Running the notebook sequentially performs: PDF parsing → chunking → embedding → FAISS index build → LangGraph agent launch. Once embeddings are stored, the entire pipeline runs fully offline, ensuring both data privacy and reproducibility in low-resource environments.

To enhance reproducibility and clarity, the entire methodology has been structured as a modular pipeline, where each stage performs a clearly defined function. From PDF parsing and chunk-based segmentation to vector embedding, FAISS indexing, LangGraph routing, passage grading, and citation-grounded generation, ev-

ery component is selected to balance performance with transparency. This architecture not only ensures that all operations remain fully offline and privacy, preserving, but also facilitates replication across different clinical or research settings. Figure 1 provides a visual reference to support comprehension of the full process.

3. Results and Discussion

The prototype is exercised through a set of representative pilot interactions embedded in the development notebook. Each user query triggered the complete RAG pipeline: the question is routed to the ACT corpus, relevant passages are retrieved from the `FAISS` index, graded for topical fit, and filtered before being passed to a locally served `llama3` model via `ChatOllama`, which returned a citation-backed answer.

Table 2 presents user prompts and representative excerpts from the replies; inline citations are omitted here for brevity, although the system includes them in actual output.

Manual inspection confirms that replies consistently reference core ACT constructs, including acceptance, values clarification, and psychological flexibility, without veering into prescriptive or unsafe medical advice. Each factual claim is grounded in a source passage, and the validation node re-queries FAISS to ensure all citations match retrieved content. In all pilot interactions, citations resolved correctly, suggesting that the retrieval–grading mechanism effectively constrains hallucinations at this early stage.

From a user experience perspective, responses are generated in real time on CPU-only hardware and use clear, empathetic language. The system successfully addresses open-ended emotional queries (e.g., "Why do I feel overwhelmed?") while maintaining ACT consistency, indicating feasibility for deployment in privacy-sensitive, patient-facing scenarios.

Current evaluation remains qualitative and limited to ACT materials; integration of oncology-specific psychoeducational content and conditional web search—already scaffolded in the code via the `tavily_search` interface, are planned next.

A structured concordance study involving licensed psycho-oncologists is underway to assess answer quality in terms of factual accuracy, therapeutic tone, and clinical adequacy. Quantitative metrics such as retrieval precision, citation overlap, and response time will also be reported once a broader question set is available.

Overall, these preliminary results demonstrate that a locally hosted RAG pipeline can deliver transparent, ACT-grounded support for individuals experiencing cancer-related distress. The modular architecture supports incremental expansion of both the corpus and the evaluation framework, laying the foundation for future clinical validation.

While this work focused on establishing a proof of concept, we recognize that ethical safeguards are essential before real-world deployment. Future iterations will incorporate critical features such as crisis escalation protocols, automated disclaimers, and pathways for optional clinical supervision. These measures are necessary to ensure the responsible use of the system and to support patient safety, especially in emotionally vulnerable scenarios.

Also, future iterations of the system will include testing with a larger and more diverse dataset, incorporating broader user characteristics and use-case contexts to assess generalizability and robustness.

4. Conclusion

This work introduces a privacy-preserving, end-to-end RAG chatbot that delivers ACT guidance tailored to the psycho-oncology context. By embedding seven authoritative ACT manuals into a `FAISS` index and orchestrating retrieval, grading, and generation via `LangChain` and `LangGraph`, the prototype demonstrates that high-quality, citation-backed responses can be generated locally using an open-source `llama3` model. Once embeddings are created, the system operates entirely offline—an essential design choice for mental-health applications where confidentiality, data locality, and cost control are critical.

From a methodological perspective, the contribution is threefold. First, it outlines a lightweight, replicable pipeline, PDF parsing, overlap-free chunking, and MiniLM-based embedding, that runs on CPU-only hardware. Second, it shows how LangGraph validation nodes can enforce strict source fidelity, effectively preventing hallucinated citations in the tested interactions. Third, it offers an extensible scaffold: the codebase already includes a Tavily search module for future integration of dynamic web evidence, and its modular architecture allows seamless expansion to oncology-specific psycho-education or additional psychotherapeutic frameworks.

Qualitative results indicate that the chatbot produces concise, empathetic responses aligned with core ACT principles while avoiding prescriptive or clinically unsafe advice—an encouraging sign of therapeutic relevance. However, this evaluation remains preliminary; larger query sets, quantitative retrieval metrics, and structured expert review by psycho-oncologists are essential before considering patient-facing deployment. Any future rollout must also address ethical safeguards, including disclaimers, crisis-response escalation, and real-time monitoring.

In summary, the project illustrates a viable pathway toward evidence-grounded conversational agents that augment psychosocial care for individuals affected by cancer. By prioritizing on-device execution and reproducibility, it reduces technical and regulatory barriers for institutions seeking to explore AI-assisted mental-health support. Ongoing work, expanding the corpus, activating conditional retrieval, and conducting expert concordance studies, will advance the prototype from proof of concept to a clinically validated, deployable system.

Table 2. Representative pilot exchanges with the prototype

User prompt	Prototype reply (excerpt)
How does a cancer diagnosis affect mental health?	A diagnosis often triggers anxiety, sadness, and uncertainty. ACT invites acknowledging these emotions while taking small, value-guided steps to maintain psychological flexibility.
How can I learn to live with grief after treatment?	Grief is a natural response. ACT recommends accepting difficult feelings without avoidance and committing to actions that honour personal values.
What does acceptance really mean in ACT?	Acceptance refers to making space for unpleasant thoughts and emotions instead of fighting them, while continuing to pursue meaningful goals.
Why do I feel emotionally overwhelmed even after being cured?	Recovery does not always bring immediate emotional closure. ACT normalises lingering discomfort and encourages value-based engagement despite inner struggle.

Ethics Statement

This study did not involve human participants or animals and therefore did not require ethical approval.

CRediT authorship contribution statement

Samara Acosta-Jiménez: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Miguel M. Mendoza-Mendoza:** Methodology, Formal analysis, Validation, Writing – review & editing. **Gerardo N. Rivera-Rojas:** Methodology, Investigation, Resources, Writing – review & editing. **Jorge I. Galván-Tejada:** Validation, Resources, Writing – review & editing. **José M. Celaya-Padilla:** Supervision, Writing – review & editing. **Carlos E. Galván-Tejada:** Conceptualization, Supervision, Project administration, Writing – review & editing.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Grammarly to assist with grammar correction and language clarity. No generative AI tools were used. All scientific content, data analyses, figures, and tables are the sole work of the authors, and all final edits were performed and verified by the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] World Health Organization, "Global cancer burden growing, amidst mounting need for services," News Release, Geneva, 2024. [Online]. Available: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024. doi: 10.3322/caac.21834 .

[3] A. J. Mitchell, M. Chan, H. Bhatti, M. Halton, L. Grassi, C. Johansen, and N. Meader, "Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: a meta-analysis of 94 interview-based studies," *The Lancet Oncology*, vol. 12, no. 2, pp. 160–174, 2011. doi: 10.1016/S1470-2045(11)70002-X .

[4] A. Pitman, S. Suleman, N. Hyde, and A. Hodgkiss, "Depression and anxiety in patients with cancer," *BMJ*, vol. 361, p. k1415, 2018. doi: 10.1136/bmj.k1415 .

[5] M. Heinrich, L. Hofmann, H. Baurecht, P. M. Kreuzer, H. Knüttel, M. F. Leitzmann, and C. Seliger, "Suicide risk and mortality among patients with cancer," *Nature Medicine*, vol. 28, no. 4, pp. 852–859, 2022. doi: 10.1038/s41591-022-01745-y .

[6] J. C. Holland, B. Andersen, W. S. Breitbart, L. O. Buchmann, B. Compas, T. L. DeShields, M. M. Dudley, S. Fleishman, C. D. Fulcher, D. B. Greenberg *et al.*, "Distress management," *Journal of the National Comprehensive Cancer Network*, vol. 11, no. 2, pp. 190–209, 2013. doi: 10.6004/jnccn.2013.0027 .

[7] F. I. Fawzy, "Psychosocial interventions for patients with cancer: what works and what doesn't," *European Journal of Cancer*, vol. 35, no. 11, pp. 1559–1564, 1999. doi: 10.1016/S0959-8049(99)00191-4 .

[8] S. L. Ehlers, K. Davis, S. M. Bluethmann, L. M. Quintiliani, J. Kendall, R. M. Ratwani, M. A. Diefenbach, and K. D. Graves, "Screening for psychosocial distress among patients with cancer: implications for clinical practice, healthcare policy, and dissemination to enhance cancer survivorship," *Translational Behavioral Medicine*, vol. 9, no. 2, pp. 282–291, 2019. doi: 10.1093/tbm/iby119 .

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[10] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial," *JMIR Mental Health*, vol. 4, no. 2, p. e7785, 2017. doi: 10.2196/mental.7785 .

[11] M. Luciano and K. Wilson, *Terapia de aceptación y compromiso: Un tratamiento conductual orientado a los valores*. Madrid: Pirámide, 2002.

[12] J. I. Cruz Gaitán, R. O. Sánchez, and P. L. Gutiérrez, *Duelo: Tratamiento basado en la Terapia de Aceptación y Compromiso*. Ciudad de México, Mexico: Editorial El Manual Moderno, 2017.

[13] S. C. Hayes, K. D. Strosahl, and K. G. Wilson, *Acceptance and Commitment Therapy: An Experiential Approach to Behavior Change*. New York, NY, USA: Guilford Press, 1999.

[14] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021. doi: 10.1109/TBDATA.2019.2921572 .

[15] Meta AI, "Meta Llama 3 — technical report," Meta AI, Tech. Rep. arXiv:2404.14219, 2024. [Online]. Available: https://arxiv.org/abs/2404.14219

[16] LangChain Developers. (2025) ChatOllama integration guide. [Online]. Available: https://python.langchain.com/docs/integrations/chat/ollama