

# Psychmet- Measurement Foundational Competencies ChatBot

Henry Makinde<sup>1</sup>, Hope Adegoke<sup>1</sup>, Mubarak Mojoyinola<sup>2</sup>

<sup>1</sup>Educational Research Methodology, University of North Carolina, Greensboro

<sup>2</sup>Psychological and Quantitative Foundations, University of Iowa, Iowa City

<sup>1</sup>hsmakinde@uncg.edu, <sup>1</sup>hoadeoke@uncg.edu, <sup>2</sup>mubarak-mojoyinola@uiowa.edu

## Abstract

We introduce PsychMet, a domain-grounded chatbot that uses a GPT-4.1 conversational model with retrieval-augmented generation over a curated psychometrics corpus (with emphasis on IRT and NCME competencies). Using the RAGAS framework on a 30-question set, PsychMet achieved Overall = 0.539, with strengths in Answer Correctness (0.810) and Context Recall (0.671), moderate Faithfulness (0.588), and weaknesses in Answer Relevancy (0.284), Context Precision (0.425), and Context Relevancy (0.474). The pattern suggests retrieval breadth outpacing specificity. We outline targeted fixes, such as hybrid sparse+dense retrieval with light filtering and question-first prompting, to tighten focus without sacrificing coverage. PsychMet is accurate and transparently sourced for exploratory learning; with retrieval tightening and answer scoping, it can better support time-bound professional workflows.

## 1 Introduction

Recently, the National Council on Measurement in Education (NCME) Presidential Task Force developed a comprehensive framework identifying essential knowledge, skills, and abilities for educational measurement professionals. The task force identified three competency domains: Communication and Collaboration; Technical, Statistical, and Computational; and Educational Measurement Competencies (Ackerman et al., 2024). Given these competencies' interdisciplinary nature, single graduate programs cannot provide comprehensive training across all domains. However, graduate education serves as the starting point for lifelong learning and continued development (Ackerman et al., 2024; Middlebrook et al., 2024; Rupp, 2024).

Given the breadth of the task force framework, individual professionals face challenges in tracking and prioritizing all competency areas. The interdisciplinary nature of educational measurement

means professionals possess diverse backgrounds in psychology, statistics, education, and computer science, requiring different developmental focus areas. Some need technical skills strengthening, while others require deeper understanding of cultural contexts or enhanced communication abilities. Therefore, there is critical need for a system that synthesizes all recommendations, systematically assesses individual competency levels, identifies improvement areas, and develops personalized learning plans with tailored resources addressing each professional's unique developmental needs.

Researchers have identified chatbot technology as capable of enhancing lifelong self-regulated learning through customized guidance and feedback (Guan et al., 2025). Chatbots use natural language processing for human-like conversation, providing educational services including homework assistance (Labadze et al., 2023), academic support (Roca et al., 2024), personalized learning (Park et al., 2024), and professional training (Terblanche and Tau, 2024).

With the demonstrated capabilities of chatbots in educational contexts, this technology can support measurement professionals in developing competencies across foundational domains. To address this need, this study aims to develop PsychMet, a large language model (LLM)-based chatbot enhanced with Retrieval Augmented Generation (RAG) technology. This system is designed to provide measurement professionals from diverse educational and training backgrounds with personalized opportunities to develop essential skills required for career success.

## 2 Related Works

**Educational Chatbots:** AI chatbots are widely used in education for Q&A, tutoring, and study support, with reviews and studies showing gains in engagement and learning across K–12, higher

ed, and professional settings (Labadze et al., 2023; Roca et al., 2024). Adaptive dialogue systems personalize feedback to learner models and improve outcomes (Park et al., 2024), while chatbot coaches can scaffold onboarding and goal attainment in the workplace (Terblanche and Tau, 2024). Collectively, these findings support chatbots as effective tools for personalized learning in specialized domains.

**Retrieval-Augmented Generation (RAG):** RAG augments LLMs with an external retriever that fetches relevant passages at query time, improving factuality and contextual relevance over relying on static pretraining (Lewis et al., 2021). In medicine, adding PubMed retrieval yielded accuracy gains comparable to moving from GPT-3.5 to GPT-4 (Xiong et al., 2024). Practically, documents are embedded into a vector store; similarity search returns pertinent chunks that are injected into the prompt, reducing hallucinations and aligning answers with current literature. RAG has proven effective for knowledge-intensive QA and dialogue and underpins PsychMet to ensure fluent, evidence-based responses to psychometrics questions.

## 3 Method

### 3.1 System Architecture

PsychMet is a modular conversational RAG system for educational measurement. It comprises (1) a large language model for Natural Language Understanding (NLU)/Natural Language Generation (NLG), (2) a document-processing pipeline and Facebook AI Similarity Search (FAISS) vector store for retrieval, (3) a retrieval chain that injects relevant context into prompts, and (4) a web UI for interactive use. On each query, the system first retrieves pertinent chunks from its corpus, then generates a source-grounded answer. Under the hood, GPT-4.1 (via LangChain) runs a Conversational RAG chain: queries are embedded with OpenAI embeddings, matched by cosine similarity against the FAISS index (top-k), and the retrieved snippets (with metadata) plus dialogue history are fed into the LLM. This keeps responses tied to the knowledge base, reduces hallucination, and enables cited answers (Lewis et al., 2021).

### 3.2 Document Processing and Retrieval

The knowledge base documents (described in the next section) are ingested through a custom document processing pipeline. PDF and text sources are

automatically chunked into smaller passages. Each chunk is vectorized by the embedding model and stored in the FAISS database along with metadata (such as source title, section, page) for traceability. This allows the system to later provide source attributions for any text used in an answer. We set the retrieval to return k=4 top chunks per query as a baseline, and tune similarity thresholds to optimize precision (filtering out marginally relevant contexts). The retrieval component emphasizes semantic search; for example, a question about “item difficulty parameters in LLTM” will retrieve the specific section in an IRT textbook covering the Linear Logistic Test Model, rather than just keyword matches. By using dense retrieval, PsychMet can handle synonyms or rephrasings robustly, which is important given the varied terminology in psychometrics.

### 3.3 LLM Integration and Prompting

PsychMet uses LangChain’s ConversationalRetrievalChain to orchestrate the interaction between the retriever and the GPT-4.1 model. This chain automatically takes the retrieved documents and the conversation history, and formats a prompt that instructs the LLM to “answer the user’s question using the provided context and cite sources.” We engineered the prompt to explicitly encourage evidence-based answers and to prefer direct quotes or specific details from the retrieved text when relevant. Memory management is handled by maintaining recent dialogue turns in the prompt so that follow-up questions are understood in context. If a user asks a follow-up like “What about 2PL models?”, the system remembers prior conversation about IRT models and retrieves accordingly. The OpenAI GPT-4.1 API (via langchain-openai) is then called with this fully constructed prompt, and the model’s completion is returned as the chatbot’s answer. Typically, the model responds with an answer that interweaves explanation with short citations (by document title or author) referencing the provided context, yielding accurate and transparent responses. (OpenAI, 2025)

### 3.4 User Interface and Deployment

The front end is a Streamlit web app that provides a chat UI with cited answers and session state to persist conversations. Deployment on Streamlit Cloud offers on-demand browser access with minimal setup. The modular design supports quick updates, adding documents or tuning retrieval with-

out system overhauls, combining a robust RAG backend with a user-friendly interface for a reliable, pedagogically useful tool.

### 3.5 Knowledge Base

PsychMet’s curated knowledge base grounds answers in authoritative psychometrics literature. The initial corpus centers on Item Response Theory (IRT) and related concepts.

To align guidance with professional standards, it also integrates the NCME Foundational Competencies framework (Ackerman et al., 2024). All sources are vetted to ensure reliability. PsychMet delivers layered explanations from quick definitions to deeper, citation-backed theory (including equations) as needed and tracks discussed topics to shape a personalized learning path, proactively suggesting related areas (e.g., model fit, DIF). The result is an interactive tutor experience that guides users toward relevant study resources, which remains grounded in respected field literature (Ackerman et al., 2024).

### 3.6 Evaluation Framework

To assess PsychMet’s performance, we adopted the Retrieval-Augmented Generation Assessment (RAGAS) framework proposed by Es et al., 2024, which is specifically designed for evaluating RAG-based systems across several dimensions of answer quality. RAGAS provides a set of metrics that cover both the retrieval component and the generation component of the system. Each metric we used is scored on a 0 to 1 scale (with higher values indicating better performance). The metrics are defined as follows:

**Context Precision:** The proportion of retrieved context documents that are relevant to the query.

**Context Recall:** The extent to which the retrieved context covers the ground truth information needed to answer the question.

**Context Relevancy:** An overall judgment of the quality and relevance of the retrieved documents.

**Faithfulness:** The degree to which the generated answer stays true to (is grounded in) the retrieved context, without introducing unsupported content.

**Answer Relevancy:** How well the answer addresses the user’s question.

**Answer Similarity:** The semantic similarity of the generated answer to the expected ground truth answer.

**Answer Correctness:** The factual accuracy of the answer compared to expert-provided correct

answers.

Using this framework, we created an evaluation dataset of representative questions that a psychometrics professional might ask. We drew 30 questions from the domain-specific materials in the knowledge base. Each question was paired with a ground truth answer compiled from the authoritative sources to serve as the reference for evaluating Answer Similarity and Correctness. PsychMet was run on each question, with no prior exposure to these specific formulated queries, and its answers were collected for analysis. For the retrieval metrics, we compared the documents fetched by PsychMet to the known relevant sections (calculating precision/recall), and for generation metrics, we utilized a combination of automated checks and LLM-based evaluators to score faithfulness, relevancy, similarity, and correctness. (Es et al., 2024)

## 4 Results

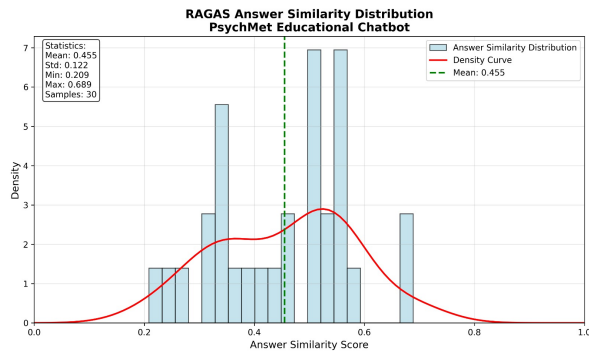
We evaluated PsychMet on a 30-question set using the RAGAS framework (Es et al., 2024). The run produced an overall score of 0.539, with component scores: Answer Correctness = 0.810, Faithfulness = 0.588, Answer Relevancy = 0.284, Answer Similarity = 0.455, Context Recall = 0.671, Context Precision = 0.425, and Context Relevancy = 0.474 (Table 1).

As visualized in the radar plot (Figure 1b), the system’s strengths are factual accuracy and evidence coverage: Answer Correctness sits in the “good–excellent” band, and Context Recall is solidly “good,” indicating that the retriever typically surfaces most of the needed material. Faithfulness is moderate, suggesting that answers are often grounded in the retrieved passages, though not uniformly so. In contrast, the weakest dimension is Answer Relevancy, which falls into the “needs improvement” band; responses frequently include background or adjacent details that do not directly answer the prompt. The retrieval layer contributes to this behavior: Context Precision and Context Relevancy are only moderate, meaning the generator receives a mix of pertinent and loosely related passages, which dilutes grounding and nudges the model toward tangents.

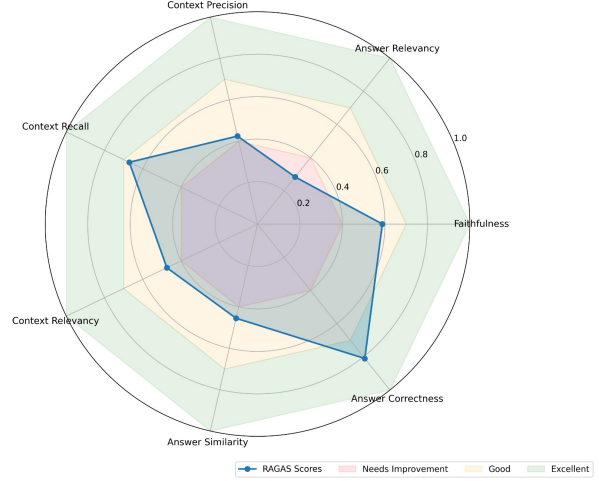
The Answer Similarity distribution (Figure 1a) centers at 0.455 with SD = 0.122 (min = 0.209, max = 0.689; n = 30). The narrow spread indicates consistent semantic alignment with reference answers across items, even when responses drift in focus;

Table 1: RAGAS Performance Metrics on 30 Examples (generated 2025-09-13).

| Metric             | Score | Interpretation                                   |
|--------------------|-------|--------------------------------------------------|
| Overall Score      | 0.539 | Needs optimization                               |
| Faithfulness       | 0.588 | Grounding is moderate; some unsupported content  |
| Answer Relevancy   | 0.284 | Answers often stray from the question focus      |
| Context Precision  | 0.425 | Retrieval includes non-relevant chunks           |
| Context Recall     | 0.671 | Most needed evidence is retrieved                |
| Context Relevancy  | 0.474 | Mixed quality of retrieved contexts              |
| Answer Similarity  | 0.455 | Semantic alignment with ground truth is moderate |
| Answer Correctness | 0.810 | Generally factually correct                      |



(a) Answer Similarity Distribution



(b) RAGAS Performance Radar Chart

there are few extreme failures or spikes. Together with the radar profile, this pattern portrays an assistant that answers correctly in substance, while it is longer and less target-specific than desired.

These results suggest two primary levers for improvement: (a) tightening retrieval to raise precision and context relevancy (thereby supporting higher faithfulness), and (b) constraining generation to produce direct, question-first answers, followed by brief, cited support. Because correctness and recall are already strong, targeted adjustments to retrieval filters and prompting should raise relevancy without sacrificing factuality or coverage.

## 5 Discussion

The 30-item evaluation portrays PsychMet as a knowledgeable but occasionally unfocused chatbot. Scores for Answer Correctness and Context Recall indicate that the system typically retrieves the right material and states core facts accurately. Where performance slips is Answer Relevancy, with Faithfulness hovering at a middle level: answers are usually grounded but sometimes drift, reflecting a familiar RAG trade-off in which recall outpaces precision (Lewis et al., 2021; Es et al., 2024). In

practice, the retriever sometimes admits loosely related passages alongside highly relevant ones; the generator then weaves a broader narrative before delivering the specific claim users expect. This is not a knowledge deficit, but rather a selection and scoping problem.

From a learning perspective, that profile has two faces. For exploration and consolidation, breadth can be helpful: contextual details connect ideas across NCME domains and support self-regulated learning with citations for follow-up (Ackerman et al., 2024; Guan et al., 2025). For time-bound settings like item reviews or committee meetings, the same expansiveness becomes friction. Stakeholders need concise, question-first answers with just enough evidence to verify and proceed. Aligning output style to intent therefore matters as much as raw accuracy. Importantly, the consistency of Answer Similarity across items suggests a stable semantic core; with more selective evidence and stricter prompting, that core can be surfaced earlier and more directly without sacrificing correctness.

These patterns also map neatly onto the NCME competency framework. PsychMet already serves the Educational Measurement and Techni-

cal/Statistical/Computational domains by retrieving authoritative content and explaining it accurately. Improving answer focus thus advances both system metrics and the very competencies the tool seeks to support (Ackerman et al., 2024; Roca et al., 2024; Park et al., 2024).

## 6 Recommendation

Two coordinated adjustments can improve performance on the broader set while preserving strengths. First, raise retrieval specificity before adding context. A hybrid retriever that fuses BM25 and dense results with reciprocal-rank fusion, applies MMR to reduce redundancy, and uses light metadata filters (same section/page/topic) will pass fewer, cleaner passages to the generator, which in turn improves Context Precision and Context Relevancy; because RAG faithfulness is evaluated against retrieved context, cleaner evidence typically yields higher Faithfulness as well (Cormack et al., 2009; Carbonell and Goldstein, 1998; Kuzi et al., 2020; Bruch et al., 2023; Es et al., 2024). Mid-length chunks (350–500 tokens) with small overlaps and preserved headings typically stabilize topicality. Second, constrain generation to answer before elaborating. A direct-answer-first instruction (one–two sentences that explicitly answer the prompt, followed by at most two cited supports) curbs digressions; a lightweight self-check that asks whether the draft directly answers the question and revises if not helps maintain focus.

Beyond mechanics, two product decisions will potentially add to these gains. Broaden and tag the corpus beyond IRT (e.g., CTT, equating, reporting, validity) and label sources to NCME competency facets so retrieval can privilege material aligned to the user’s purpose. Keep the 30-item set as a standing benchmark, add SME ratings alongside RAGAS to triangulate LLM-based judgments, and track a small set of iteration KPIs (e.g., like a +0.10 increase in Answer Relevancy, +0.08 in Context Precision) so the radar profile tightens around a high-precision, high-faithfulness center (Es et al., 2024; Guan et al., 2025).

In short, PsychMet does not need more facts to perform better on harder queries; it needs to show the right facts sooner and more selectively. With tighter retrieval and question-first prompting, the system should retain its strong correctness and coverage while delivering the concise, cited responses that practitioners expect.

## 7 Limitations

Our findings should be read with several constraints in mind. First, the 30-item evaluation is modest and author-curated, so item phrasing and topical mix may not reflect the full diversity of real queries, limiting generalizability. Second, the knowledge base is IRT-heavy; performance on underrepresented areas (e.g., CTT, equating, reporting, validity) is less certain. Third, several RAGAS dimensions rely on LLM-based judgments, introducing evaluator variance absent parallel SME ratings (Es et al., 2024). Fourth, we report a single configuration without ablations or baselines (e.g., non-RAG LLMs or search-only), so causal attributions for low answer relevancy versus retrieval precision remain tentative. Fifth, we did not measure user-centered outcomes (learning gains, time-on-task, usefulness/trust) or operational factors (latency, cost), and reproducibility is affected by proprietary API updates. Finally, the corpus is static and unaudited for fairness; without routine updates, drift may erode retrieval quality, and differential performance remains unassessed (Lewis et al., 2021). Nonetheless, this represents an initial implementation of the system, with planned knowledge base expansion expected to address current coverage limitations and enhance overall performance in future iterations. The current implementation is accessible at <https://psychmet-chatbot.streamlit.app> for demonstration and evaluation purposes.

## References

- Terry A Ackerman, Deborah L Bandalos, Derek C Briggs, Howard T Everson, Andrew D Ho, Susan M Lottridge, and Stefanie A Wind. 2024. Foundational competencies in educational measurement. *Educational Measurement: Issues and Practice*, 43(3):7–17.
- Sebastian Bruch, Siyu Gai, and Amir Ingber. 2023. [An analysis of fusion functions for hybrid retrieval](#). *ACM Transactions on Information Systems*, 42(1):1–35.
- Jaime G. Carbonell and Jade Goldstein. 1998. [The use of MMR, diversity-based reranking for reordering documents and producing summaries](#). In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336. ACM.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR*

- Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.
- Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. [RAGAS: Automated evaluation of retrieval augmented generation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, St. Julian’s, Malta. Association for Computational Linguistics.
- Rui Guan, Mladen Raković, Guanliang Chen, and Dragan Gašević. 2025. How educational chatbots support self-regulated learning? a systematic review of the literature. *Education and Information Technologies*, 30(4):4493–4518.
- Saar Kuzi, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. [Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach](#). *Preprint*, arXiv:2010.01195.
- Lela Labadze, Maka Grigolia, and Lela Machaidze. 2023. Role of ai chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education*, 20(1):56.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). *Preprint*, arXiv:2005.11401.
- Kristen Middlebrook, Laura S Hamilton, and Megan E Walker. 2024. Commentary: How research and testing companies can support early-career measurement professionals. *Educational Measurement: Issues and Practice*, 43(3):45–48.
- OpenAI. 2025. Introducing gpt-4.1 in the api. <https://openai.com/index/gpt-4-1/>. Accessed: 2025-09-15.
- Minji Park, Soobin Kim, Seunghyun Lee, Sungho Kwon, and Kyungmin Kim. 2024. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–10.
- Maria de los Remedios Roca, Mou Meng Chan, Antonio Garcia-Cabot, Eva Garcia-Lopez, and Hector Amado-Salvatierra. 2024. The impact of a chatbot working as an assistant in a course for supporting student learning and engagement. *Computer Applications in Engineering Education*, 32(5):e22750.
- Andre A Rupp. 2024. Commentary: Modernizing educational assessment training for changing job markets. *Educational Measurement: Issues and Practice*, 43(3):33–38.
- Nico HD Terblanche and Thabo Tau. 2024. Exploring the use of a goal-attainment, artificial intelligence (ai) chatbot coach to support first-time graduate employees. *Industry and Higher Education*, page 09504222241287090.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. [Benchmarking retrieval-augmented generation for medicine](#). *Preprint*, arXiv:2402.13178.