

Assessing AI-Assisted Search Answer Integrity on Digital Banking Websites: A Comparative Audit of US and Pakistani Banks



MUHAMMAD MOIN UDDIN
01-322242-012

Supervisor: Ms. Nadia Kalsoom

Masters of Business Administration

Department of Human Resource and Management
Bahria University, Islamabad

May 2026

Abstract

AI assistants are now used to answer everyday questions, including about banking, yet their accuracy in this domain is largely untested. A wrong fee or interest rate is not academic; it can cost the consumer real money. Whether accuracy holds across markets, or whether phrasing changes the answer, is also unknown.

The design is a cross sectional comparative audit of the top five banks by total assets in the United States and Pakistan. Each was queried about four consumer-banking facts (two fees, a savings rate and a 24/7 helpline) under a structured source-anchored prompt and a short naturalistic one, each repeated twice for 160 LLM responses. Ground truth was built manually from the official websites, and each response was classified into five categories by an LLM-as-judge validated against two human annotators.

Even with web browsing on, the correctness rate across the 160 responses is 51.3% and the hallucination rate is 31.9%. The country difference is significant ($\chi^2 = 16.31$, $p = 0.0026$, Cramér's $V = 0.32$): U.S. banks are Correct two thirds of the time, Pakistani banks just over one-third. The naturalistic prompt scores higher than the structured one (57.5% vs. 45.0%), but a paired McNemar's on 40 pairs is not significant, so the claim is directional. Run-to-run consistency is 83-85%.

The study contributes a replicable framework, the five-category taxonomy, hybrid automated/human verification and paired prompt-formality design, for evaluating AI accuracy wherever a verifiable reference is available, including fields beyond banking. It also sheds light on how the way banks construct information shapes what AI assistants report.

Acknowledgments

I would like to gratefully acknowledge the guidance and constructive feedback which she provided me during the course of this research and for the continuous academic support being provided by Ms Nadia Kalsoom, my thesis supervisor. Her observations and insightful feedback and encouragement, have been instrumental in the direction, structure and quality of this study.

I would also like to express my gratitude to the faculty members of the Department of Human Resource & Management, Business School, Bahria University Islamabad for having provided a challenging academic environment and for their insightful inputs during coursework and discussions which helped to develop this research.

I would also like to express my gratitude to the two independent raters that agreed to take part in the inter-rater reliability check anonymously. They independently judged a stratified sample of the AI answers based on the same five-point taxonomy employed by the automated judge, and assigned to the same 5 categories as the automated judge. It was their careful work which included a second coding pass during the calibration step which allowed this step of verification in this study. Due to the anonymous nature of the raters, they are only called Rater 1 and Rater 2 throughout the thesis.

Also, I would like to thank Bahria University, Islamabad, for the institutional support and academic resources, which helped me in the successful completion of this thesis. Facilities, research atmosphere and administrative support played a key role in the execution of this work in a timely and orderly fashion

MOIN UDDIN
Bahria University
E-8, Islamabad, Pakistan

May 2026

Contents

Abstract	i
Acknowledgments	ii
Acronyms and Abbreviations	ix
1 Introduction	1
1.1 Background	1
1.2 Problem Statement	2
1.3 Aim and Objectives	3
1.4 Research Questions	4
1.5 Significance of the Study	4
1.5.1 Theoretical Contribution	5
1.5.2 Methodological Contribution	5
1.5.3 Practical Contribution	5
1.5.4 Cross-Country Insights	5
1.6 Scope and Delimitations	5
1.7 Key Term Definitions	6
1.8 Structure of the Thesis	6
1.9 Chapter Summary	7
2 Literature Review	8
2.1 Introduction	8
2.2 Information Retrieval: Principles and History	8
2.2.1 Traditional Information Retrieval	8
2.2.2 Search Engine Optimization and Information Architecture	9
2.3 Large Language Models	9
2.3.1 Origins of Language Modeling	9
2.3.2 Capabilities and Applications	9
2.3.3 Limitations and Failure Modes	10
2.4 Retrieval-Augmented Generation	10
2.4.1 Conceptual Framework	10
2.4.2 Developing RAG Systems	11
2.4.3 Production Systems of RAG	11
2.5 Hallucination in Large Language Models	11
2.5.1 Definition of Hallucination	11
2.5.2 Detection and Measurement of Hallucinations	12
2.5.3 Strategies for the Mitigation of Hallucination	12

2.6	Evaluation Frameworks for AI Systems	12
2.6.1	Classical Evaluation Metrics for Natural Language Processing	12
2.6.2	Evaluation of RAG Systems	13
2.6.3	Human Evaluation, Inter-Rater Reliability, and LLM-as-Judge	13
2.6.4	Benchmark Datasets and Evaluation Campaigns	14
2.7	Verifiability and Citation in Generative Search Engines	14
2.7.1	The Verifiability Challenge	14
2.7.2	Citation Accuracy and Source Attribution	14
2.7.3	The Illusion of Verifiability and Its Consequences for User Trust	15
2.8	Non-Determinism in Large Language Models' Behavior	15
2.8.1	Sources of Variability	15
2.8.2	Implications for Evaluation	15
2.8.3	Consistency as an Evaluation Dimension	16
2.9	AI Applications in Financial Services	16
2.9.1	Financial Services Context	16
2.9.2	Consumer Information Needs	16
2.9.3	Banking Website Information Architecture	17
2.10	Cross-Country Considerations in AI Evaluation	17
2.10.1	Geographic Bias in AI Systems	17
2.10.2	Comparison of Developed and Emerging Markets	18
2.10.3	Generalizability of Evaluation Findings	18
2.11	Synthesis and Research Gaps	18
2.11.1	Identified Gaps	18
2.12	Conceptual Framework	20
2.13	Chapter Summary	21
3	Methodology	22
3.1	Research Design	22
3.2	Sampling Strategy	23
3.2.1	Bank Selection	23
3.2.2	Question Selection	23
3.3	Ground Truth Construction	23
3.3.1	Data Collection Process	24
3.3.2	Organization of the Reference Dataset	24
3.4	LLM Querying Protocol	25
3.4.1	Query Formulation	25
3.4.2	Repeated Sampling	25
3.5	Response Classification	26
3.6	Assessment of Classification Reliability	28
3.6.1	Classification Procedures	28
3.7	Data Analysis	29
3.7.1	Descriptive Statistics (RQ1)	29
3.7.2	Cross-Country Comparison (RQ2)	29
3.7.3	Prompt-Formality Effect (RQ3)	29
3.7.4	Within-Prompt Consistency (RQ4)	30
3.8	Ethical Considerations	30
3.9	Methodological Limitations	30
3.10	Summary	31

4	Findings and Analysis	32
4.1	Overview	32
4.1.1	Sample and Coverage	32
4.2	Descriptive Correctness (RQ1)	33
4.2.1	Overall Distribution	33
4.2.2	Distribution by Country	34
4.2.3	Distribution by Question Category	35
4.2.4	Distribution by Prompt Form	35
4.3	Cross-Country Comparison (RQ2)	37
4.3.1	Omnibus Test	37
4.3.2	Cell-Level Drivers	37
4.3.3	Sub-Analysis by Question Category	38
4.4	Effect of Prompt Formality (RQ3)	38
4.4.1	Pairing	38
4.4.2	Binary McNemar Test	39
4.4.3	Full Five-Category Marginal Homogeneity	39
4.5	Within-Prompt Consistency (RQ4)	40
4.6	Diagnostic Findings	40
4.6.1	Browsing Behaviour	40
4.6.2	Citation Behaviour	41
4.6.3	Response Length	41
4.7	Notable Cases	41
4.8	Reliability of the Classification Process	42
4.9	Summary of Findings	43
5	Conclusions and Recommendations	45
5.1	Overview	45
5.2	Discussion of Findings	45
5.2.1	Overall Correctness and Persistent Hallucination Despite Browsing	45
5.2.2	A Significant Cross-Country Gap Concentrated in Outright Error	45
5.2.3	Humanized Prompts: Higher Correctness but Weaker Citations	46
5.2.4	Response Consistency Across Repeated Runs	47
5.3	Implications and Recommendations	47
5.3.1	For banks	47
5.3.2	For consumers	48
5.3.3	For AI designers	49
5.4	Contribution of the Study	49
5.4.1	Empirical Contribution	49
5.4.2	Methodological Contribution	50
5.4.3	Practical Contribution	50
5.4.4	Cross-Country Contribution	50
5.5	Limitations	50
5.6	Future Research	51
5.7	Closing	52

6	Reflections	53
6.1	Overview	53
6.2	What the Ground Truth Taught Me	53
6.3	Three Things That Surprised Me	54
6.4	What I Would Do Differently	54
6.5	Using AI to Study AI	55
6.6	What the Study Has Left Me Thinking About	55
6.7	Closing	56
A	Sample Banks and Ground Truth Reference	57
A.1	Sample Banks	57
A.2	Ground Truth Values	57
	References	61

List of Figures

1.1	Conceptual framework of AI-risk in consumer banking information retrieval and the proposed audit methodology.	3
2.1	Schematic overview of a retrieval-augmented generation (RAG) pipeline.	10
3.1	Experimental design pipeline, from ground-truth construction through LLM querying and five-category classification to the statistical analysis.	22
3.2	Decision tree for the five-category response classification taxonomy.	27
4.1	Distribution of classification categories by country (per-country %) with Wilson 95% confidence intervals.	34
4.2	Interaction plot of prompt form and country on the Correct rate.	36

List of Tables

2.1	Gap analysis mapping representative prior work to the gaps addressed by this study.	19
3.1	Question categories, the representative question used, and the rationale for inclusion.	24
3.2	Humanized query variants per question and country.	26
3.3	Response classification taxonomy.	27
4.1	Sample coverage by country, question, and prompt form.	33
4.2	Distribution of classification categories across all responses, with Wilson 95% confidence intervals.	33
4.3	Classification distribution by country.	34
4.4	Classification distribution by question category.	35
4.5	Classification distribution by prompt form.	36
4.6	Classification distribution by country and prompt form.	36
4.7	Contingency table of classification category by country.	37
4.8	Omnibus chi-square test of independence on the category $\times$ country contingency table.	37
4.9	Standardised residuals for the category $\times$ country contingency table.	38
4.10	Per-question-category chi-square sub-analysis of the country effect.	38
4.11	McNemar’s exact test on Correct versus Not Correct, paired by (bank, question).	39
4.12	Stuart-Maxwell test of marginal homogeneity across the five-category outcome.	39
4.13	Within-prompt consistency: proportion of (bank, question, prompt form) tuples in which the two runs received the same classification.	40
4.14	Citation-count and first-party citation-rate diagnostics by prompt form.	41
4.15	Response length (characters) by prompt form.	41
4.16	Pairwise Cohen’s kappa for the verification sample ($n = 72$).	43
A.1	Banks included in the sample.	57
A.2	Ground truth values per (bank, question) pair, with source URL and access date.	58

Acronyms and Abbreviations

AI	Artificial Intelligence
ATM	Automated Teller Machine
CD	Certificate of Deposit
DPR	Dense Passage Retrieval
FEVER	Fact Extraction and VERification
GEO	Generative Engine Optimization
GT	Ground Truth
HTML	HyperText Markup Language
IR	Information Retrieval
LLM	Large Language Model
NLP	Natural Language Processing
OCR	Optical Character Recognition
PDF	Portable Document Format
POSIX	Prompt Sensitivity Index
RAG	Retrieval-Augmented Generation
RAGAs	Retrieval Augmented Generation Assessment
RARR	Researching and Revising What Language Models Say
RL	Reinforcement Learning
RQ	Research Question
SEO	Search Engine Optimization
SQuAD	Stanford Question Answering Dataset
TF-IDF	Term Frequency-Inverse Document Frequency
TREC	Text Retrieval Conference
URL	Uniform Resource Locator
US/USA	United States of America