

Article

Toward Sustainable AI in Digital Financial Services: A Hybrid Data-Grounded Architecture for Reducing Generative Model Dependency

Busra Ozdenizci Kose ¹, Vedat Coskun ^{2,*} and Ulas Baysalli ³

¹ Department of Management Information Systems, Gebze Technical University, Kocaeli 41400, Türkiye; busraozdenizci@gtu.edu.tr

² Department of Computer Engineering, Istanbul Atlas University, Istanbul 34408, Türkiye

³ Konfides Information Technologies, Istanbul 34732, Türkiye; ulas.baysalli@konfides.com

* Correspondence: vedat.coskun@atlas.edu.tr

Abstract

AI assistants are increasingly integrated into digital financial services; however, relying on generative language models for response generation can introduce challenges related to response reliability, data exposure, and resource requirements. This study presents a hybrid data-grounded architecture that selectively uses generative AI while maintaining reliable natural language interaction. The proposed task-oriented system combines semantic intent matching, slot and entity extraction, reconciliation, and response routing. Instead of using the generative model to describe every response, the architecture constructs template-based responses directly from structured institutional data whenever sufficient information is available and invokes the generative language model only as a fallback. The architecture was evaluated using a fixed set of 100 realistic user queries, including diverse expressions and spelling variations, across 14 iterative system refinements while the underlying AI models remained unchanged. Response correctness was manually assessed against the underlying data and classified as correct, unanswered, or incorrect. Correct responses increased from 51% to 97%, whereas incorrect responses decreased from 33% to 0%. Simultaneously, language-model-generated responses decreased from 34% to 10%, whereas data-grounded deterministic responses increased from 66% to 90%. These findings indicate that system-level architectural refinement can substantially improve response reliability while reducing dependence on generative response production. However, the observed reduction in generative-model dependency should be interpreted as a potential resource-conscious architectural strategy rather than evidence of a measured environmental benefit.

Keywords: sustainable AI; digital financial services; generative AI; hybrid AI architecture; data-grounded response generation; task-oriented dialogue systems; trustworthy AI



Academic Editor: Cristina Raluca Gh. Popescu

Received: 21 August 2026

Revised: 20 September 2026

Accepted: 22 September 2026

Published: 28 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Artificial intelligence (AI) has become an increasingly important component of digital service environments, enabling users to access information and interact with complex systems through natural language. Recent advances in large language models (LLMs) have further expanded these capabilities by enabling flexible interpretation of user requests and fluent response generation [1,2]. However, the growing adoption of generative AI has intensified concerns regarding the computational resources required for AI development and deployment [3,4]. Green AI emphasizes that advances in model performance should

be considered together with computational efficiency and resource requirements [3,4]. As generative AI becomes embedded in everyday digital services, these considerations increasingly extend from model training to the repeated use of models during inference [4,5].

These issues are particularly relevant to task-oriented digital services. Unlike open-domain conversational systems, task-oriented AI assistants are designed to help users accomplish specific goals, such as retrieving information, locating services, or accessing transaction-related information [2,6,7]. Task-oriented dialogue systems have traditionally addressed these interactions through components such as intent recognition, slot filling, dialogue management, and response generation [2,7]. In many domain-specific applications, structured organizational data already contain the information required to answer a request. Therefore, the primary challenge is not always to generate new content but to correctly interpret the request, identify the required parameters, retrieve the appropriate information, and communicate it accurately.

The emergence of LLMs provides greater flexibility for understanding diverse natural-language expressions, but it does not eliminate the need to ground responses in reliable external information [6,7]. This requirement is especially important in digital financial services, where LLM-based assistants are increasingly being investigated for digital payments, banking, and personal financial services [8–11]. Responses in such environments should remain consistent with institutional data rather than solely relying on plausible language generation. Retrieval-augmented and grounded AI approaches have consequently emphasized the integration of language models with external information sources and controlled processing mechanisms [7,12]. However, unlike conventional retrieval-augmented generation, structured information does not always need to be passed to a generative model for final response formulation. When sufficient structured information is available, response construction may be performed through controlled, data-grounded mechanisms, while generative models are reserved for requests that cannot be adequately resolved through such processing.

This distinction introduces an important system-design question from the perspective of sustainable AI: Does every user request require generative response production? If a request can be reliably resolved after identifying the user's intent and relevant parameters and retrieving the necessary structured data, a deterministic data-grounded response may be sufficient. Then, generative processing can be reserved for requests that cannot be adequately resolved through these mechanisms. Such an approach shifts attention from model capability alone to orchestration of the overall AI system. This perspective is consistent with broader discussions of resource-conscious AI, which emphasize that computational requirements should be considered alongside system performance [3–5].

Despite these developments, limited attention has been paid to how system-level refinements influence both response reliability and the need for generative response production within a hybrid fixed architecture. This gap is particularly relevant in digital financial services, where reliable access to structured and institution-specific information is a primary operational requirement. Examining these mechanisms provides an opportunity to assess whether generative AI can be used selectively without compromising response reliability.

This study investigates a hybrid, data-grounded AI assistant developed for a digital payment environment. The proposed architecture separates language understanding from response generation by combining semantic intent matching and parameter extraction with reconciliation and response routing. When a user request can be resolved using available structured data, the response is constructed through a deterministic, data-grounded mechanism. When the request cannot be adequately resolved, a generative language model is used as a fallback. This study aims to investigate whether systematic refinement of

this hybrid architecture can improve response reliability while reducing dependence on generative language-model-based response production without replacing or retraining the underlying AI model architectures. Accordingly, this study addresses the following research questions:

RQ1. *To what extent can the iterative refinement of a hybrid data-grounded AI architecture improve the correctness of response in a task-oriented digital financial assistant?*

RQ2. *To what extent can such refinements reduce dependence on generative language-model-based response production?*

RQ3. *What design insights emerge from the iterative refinement of the hybrid data-grounded architecture?*

This study contributes a system-level perspective to the development of sustainable and trustworthy AI assistants. Rather than assuming that improved performance necessarily requires greater reliance on generative models, this study examines how structured data processing, semantic understanding, deterministic response generation, and selective generative-model use can be orchestrated within a unified architecture. In this context, reduced generative-model dependency is examined as a resource-conscious architectural strategy while maintaining response reliability as the primary operational requirement.

The remainder of this paper is organized as follows: The research methodology, hybrid system architecture, evaluation dataset, iterative refinement process, and evaluation procedure. Section 3 presents the results regarding response reliability, generative-model dependency, and the effects of the iterative refinements. Section 4 discusses the findings in relation to system-level reliability, selective generative AI use, design knowledge, and sustainable AI implications, and outlines the limitations of the study and future research directions. Finally, Section 5 concludes the paper by summarizing the main findings and contributions.

2. Materials and Methods

2.1. Research Design

This study was structured using a Design Science Research (DSR) perspective, which is appropriate for research concerned with the design, development, and evaluation of technological artifacts intended to address identified problems [13,14]. In DSR, knowledge about a problem and its potential solution is generated through the construction and evaluation of an artifact [13,14]. The DSR process proposed by Peffers et al. [15] further organizes this approach around problem identification and motivation, solution objective definition, design and development, demonstration, evaluation, and communication.

The research artifact in this study was a hybrid, data-grounded AI assistant architecture developed for a digital payment environment. The research process addressed the reliability limitations of the existing assistant through iterative refinement and artifact evaluation. Instead of replacing the underlying AI models, the refinement process focused on the surrounding system architecture, including structured knowledge, lexical resources, entity processing, reconciliation mechanisms, and response logic. This approach enabled the study to examine how changes in system-level design affected both response reliability and the use of generative response production.

The empirical evaluation followed a fixed-query iterative design with 100 realistic user queries and 14 sequential system-level refinements. The underlying semantic and generative models remained unchanged throughout the process. Table 1 summarizes how the main stages of DSR were operationalized in this study.

Table 1. Design Science Research process applied in this study.

Design Science Research Stage	Implementation in This Study
Identification of problems and motivation	Response reliability limitations were identified when the AI assistant was evaluated using realistic user queries rather than the original developer-created test set.
Objectives of the solution	To improve response reliability while reducing unnecessary dependence on generative response production.
Design and development	The development and iterative refinement of a hybrid architecture combining semantic intent matching, parameter extraction, reconciliation, structured data processing, deterministic response generation, and generative-model fallback.
Demonstration	The artifact is applied to a fixed set of 100 realistic queries representing different request types within the digital payment environment.
Evaluation	Assessment of system responses against the underlying reference data using predefined response quality categories and recording of the response generation pathway.
Iterative refinement	Sequential introduction and evaluation of 14 targeted system-level refinements using the same fixed query set.

2.2. Research Artifact: Hybrid Data-Grounded Artificial Intelligence Architecture

The research artifact is a task-oriented AI assistant designed to process natural-language requests within a digital payment environment. The proposed architecture separates language understanding from response generation and combines semantic intent matching, parameter extraction, reconciliation, response routing, deterministic data-grounded response generation, and generative language-model fallback.

Two AI models perform distinct functions within the architecture. The semantic matching component uses bge-m3, a 566-million-parameter model, to identify the user's intended task. When the preceding processing stages cannot adequately resolve the request, the generative component uses Qwen3:8B, an 8-billion-parameter language model. Importantly, the semantic model remains involved in requests handled through the deterministic pathway. Thus, the distinction between the two response pathways does not concern whether AI is used but whether the final response is constructed from structured institutional data through predefined response logic or generated by the language model.

The processing pipeline begins with semantic intent matching and parameter extraction. Semantic matching determines the task represented by the user's request, while parameter extraction identifies structured information, such as merchant category, subtype, brand, and location, required to resolve it. Entity identification may rely on dictionary-based mechanisms where semantic similarity could introduce ambiguity. For example, location names are resolved through structured lexical matching rather than semantic proximity.

Subsequently, a reconciliation mechanism evaluates the consistency between the inferred intent and extracted parameters. It can reject an apparently confident intent when structured evidence contradicts the inference or, conversely, resolve an otherwise uncertain request when the extracted parameters provide sufficient evidence. Following reconciliation, the routing mechanism determines whether the request can be resolved using the available structured information.

When sufficient information is available, the system follows the deterministic data-grounded pathway. Each supported intent is associated with a predefined response structure populated with values retrieved from the underlying institutional data. In this pathway, the generative language model is not used to make the final response. For example, for a query such as "Can I use my card at McDonald's?" the system can identify the card-

acceptance intent and the relevant brand parameter, retrieve the corresponding merchant information from the institutional data, and populate the predefined response structure without requiring the generative language model to formulate the final answer. When the request cannot be adequately resolved through the preceding stages, it is routed to the generative language model as a fallback mechanism. Figure 1 summarizes the resulting hybrid architecture.

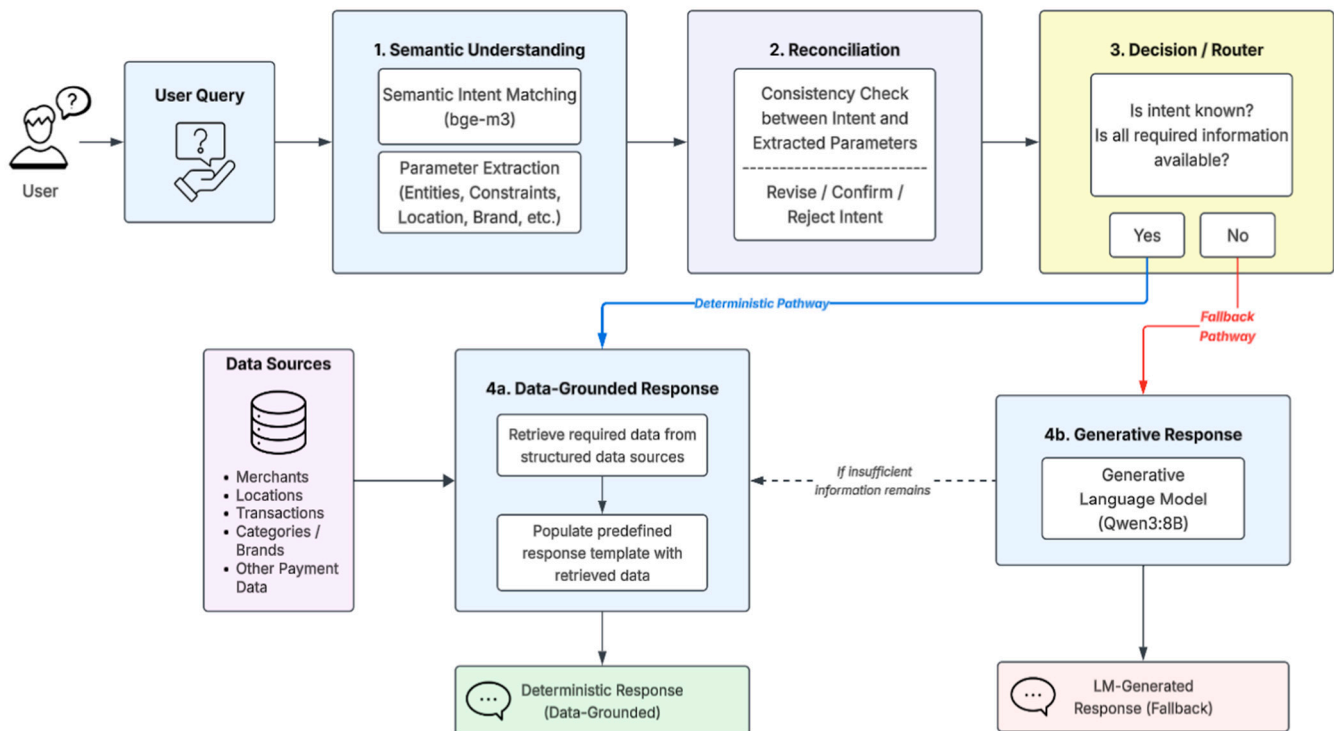


Figure 1. Hybrid data-grounded AI assistant architecture (Source: Developed by the authors).

2.3. Evaluation of the Query Set

The artifact was evaluated using a fixed set of 100 realistic user queries developed separately from the test set created by the original developer. The purpose of constructing a separate evaluation set was to assess the AI using natural user expressions rather than queries developed to match expressions already anticipated during system development. The query set was designed to cover multiple request types supported within the digital payment environment, including brand-specific card acceptance, product- and market-based merchant searches, location-related requests, unsupported payment categories, and out-of-domain or follow-up requests. Therefore, the selection criteria emphasized the coverage of the main supported request types, variation in user wording and linguistic form, and independence from expressions already anticipated in the original developer-created test set. Accordingly, the evaluation set was intended to provide a consistent and diverse test basis for iterative artifact evaluation within the target application context, rather than a statistically representative sample of the broader financial-services domain.

All 100 queries were submitted in a single initial evaluation session without modifying the system configuration. The evaluation set covered seven categories of requests: 29 queries concerning card acceptance by specific brands, 18 product-based merchant searches, 17 market searches, 13 queries concerning alcohol or unsupported payment categories, 12 location-specific requests, seven location or address requests, and four out-of-domain and follow-up requests. The distribution of the evaluation queries is presented in Table 2.

Table 2. Distribution of the evaluation query set.

Query Category	Number of Queries	Representative Query
Brand-specific acceptance of cards	29	“Can I use the card at Starbucks?”
Product-based merchant search	18	“Is there anywhere nearby that sells pizza?”
Market search	17	“Where is the nearest market that accepts the meal card?”
Alcohol and unsupported payment categories	13	“Can I buy a cocktail with this card?”
Location-specific queries	12	“Where can I eat fish in Karşıyaka?”
Location and address queries	7	“How can I get to Kahve Dünyası?”
Other queries	4	“How often is the World Cup held?”

Note: The queries were originally submitted in Turkish. Representative examples are translated into English for ease of reading. The 40 queries contained spelling errors or nonstandard expressions that were retained during the evaluation. The “Other” category includes out-of-domain and follow-up queries.

The linguistic variation was deliberately retained in the evaluation set. Of the 100 queries, 40 contained spelling errors or nonstandard written expressions. These variations were preserved rather than normalized before the initial evaluation because they represented realistic user input and allowed the system to be assessed under linguistically diverse conditions. The same fixed set of 100 queries, including these variations, was retained throughout all iterative evaluation runs. This fixed-query design enabled a direct comparison of system behavior across successive refinements while controlling for evaluation input variation.

Because the same 100 queries were used throughout the evaluation, paired analysis was used to compare baseline and final results. Response correctness was grouped as Correct and Not Correct (Unanswered or Incorrect), and the exact McNemar test was used to test the change. A 95% bootstrap confidence interval was also calculated for the change in correct responses. The same paired approach was used to compare the data-grounded deterministic and LLM-generated response pathways.

2.4. The Iterative Build-Evaluate-Refine Procedure

Following the baseline evaluation, a 14-step iterative refinement process was conducted. Three methodological rules were maintained throughout the study.

The evaluation query set was frozen after its initial construction and was not modified during subsequent refinements. This ensured that system performance changes were evaluated against the same set of user queries rather than against progressively modified or simplified evaluation inputs.

Second, each refinement targeted a specific problem identified in the previous evaluation. The modification was incorporated into the system, and the complete set of 100 queries was rerun before proceeding to the next refinement. Therefore, the refinements were introduced sequentially, with each evaluation representing the updated system configuration at that stage of the iterative process. This procedure enabled the consistent tracking of changes in system behavior across successive iterations.

Third, system responses were manually reviewed rather than relying exclusively on automated evaluation. Each response was checked against the underlying reference data to determine whether the requested information had been provided and whether the available data supported the factual information contained in the response.

The refinements focused on the system-level mechanisms surrounding the semantic and generative models rather than replacing the models themselves. These included changes to structured brand and category information, communication of previously resolved information, confidence handling, response formulations, lexical dictionaries, example expressions used for semantic matching, distinctions between unknown and unavailable information, category-processing logic, merchant subtype representation, linguistic normalization, list-response consistency, and presentation of financial responses. Table 3 summarizes the 14 refinements and the targeted system components at each step.

Table 3. Applying iterative refinements to the hybrid AI artifact.

Step	System-Level Refinement	Targeted Component
1	Structured brand information was provided to the generative model.	Structured knowledge/model context
2	Structured category information was provided to the generative model.	Structured knowledge/model context
3	Previously resolved information was communicated to the generative model.	Model context
4	Confidence levels associated with resolved information were differentiated.	Confidence handling
5	Response formulations were revised.	Response generation
6	Lexical dictionaries were enriched.	Entity and parameter processing
7	Additional query forms were added to the example expressions used for semantic matching.	Semantic intent matching
8	Unknown information was distinguished from confirmed absence.	Decision logic
9	Category-processing logic was revised to prevent product names from being incorrectly treated as merchant categories.	Category processing
10	Merchant subtypes were differentiated within the structured data.	Structured data/entity processing
11	Attached question suffixes were separated during query processing.	Query processing
12	Missing suffix and spelling normalization mechanisms were introduced.	Linguistic normalization
13	Response behavior for list-based queries was made more consistent.	Response generation
14	Visual components were incorporated into financial responses.	Response presentation

Note: Each refinement was evaluated using the same fixed set of 100 queries. The bge-m3 and Qwen3:8B models remained unchanged throughout the process.

2.5. Evaluation Criteria and Measures

Each response was manually classified into one of three mutually exclusive outcome categories:

- **Correct:** The requested information was provided, and all factual information in the response was consistent with the underlying reference data.
- **Unanswered:** The system did not provide the requested information, but did not provide incorrect information.
- **Incorrect:** The response contained information that was inconsistent with the reference data or that responded to a different request.

The final manual assessment was used as the reference evaluation. During system development, an automated evaluator was also available; however, it was not used as the sole basis for response classification. Therefore, manual verification against the underlying reference data was retained as the basis for the reported response outcomes.

The response-generation pathway was recorded for each query. Responses were classified as either data-grounded deterministic responses, in which the final response was

constructed from structured data using predefined response templates, or language-model-generated responses, in which the response text was produced by the generative model. Tracking these two pathways across successive system configurations enabled the examination of changes in generative-model dependency alongside changes in response reliability.

Additionally, response stability was examined across the iterative evaluation process. The number of changes in response outcome classification (correct, unanswered, or incorrect) across successive system configurations was recorded. These changes were then summarized descriptively according to the final response-generation pathway. This enabled the comparison of the stability of response outcomes for queries ultimately resolved through the deterministic data-grounded pathway and those remaining dependent on language-model-generated responses.

3. Results

3.1. Overall Improvement in Response Reliability

The iterative refinement process substantially improved the overall response reliability. In the baseline evaluation, 51 of the 100 queries were answered correctly, 16 remained unanswered, and 33 received incorrect responses. Following the fourteen refinement steps, the final system correctly answered 97 queries, while three queries remained unanswered. Accordingly, the proportion of correct responses increased from 51% to 97%, whereas that of incorrect responses decreased from 33% to 0%.

The improvement progressively developed across the iterative evaluation process. The number of correct responses increased by 46, while unanswered responses decreased by 13 and incorrect responses decreased by 33. Overall, the combined number of unanswered and incorrect responses declined from 49 in the baseline configuration to three in the final configuration.

The paired analysis showed that 46 queries changed from Not Correct at baseline to Correct in the final configuration, while no query changed from Correct to Not Correct. This represents a 46-percentage-point increase in correct responses (95% bootstrap CI: 36–56 percentage points), with a significant change according to the exact McNemar test ($p < 0.001$).

3.2. Reduction in Generative-Model Dependency

The iterative refinement process also produced a substantial change in the manner in which responses were generated within the hybrid architecture. In the baseline configuration, 66 of the 100 queries were handled through the data-grounded deterministic pathway, whereas 34 required language-model-generated responses. In the final configuration, the number of data-grounded deterministic responses increased to 90, whereas the number of language-model-generated responses decreased to 10. Accordingly, the proportion of queries requiring generative response production decreased from 34% to 10%, representing a reduction of 24 percentage points and a relative reduction of 70.5% in generative language model invocations.

This reduction in generative-model use occurred while the overall response correctness increased, as reported in Section 3.1. The paired analysis showed that 24 queries shifted from the LLM-generated pathway at baseline to the data-grounded deterministic pathway in the final configuration, whereas no query shifted in the opposite direction. This represents a 24-percentage-point decrease in generative-model use (95% bootstrap CI: 16–33 percentage points), with a significant change according to the exact McNemar test ($p < 0.001$). Table 4 summarizes the changes in the response-generation pathways between baseline and final configurations.

Table 4. Response generation pathways in baseline and final configurations.

Response Pathway	Baseline (%)	Final (%)	Change (pp)
Data-Grounded Deterministic Response	66.0%	90.0%	+24.0
Language-model-generated response	34.0%	10.0%	−24.0

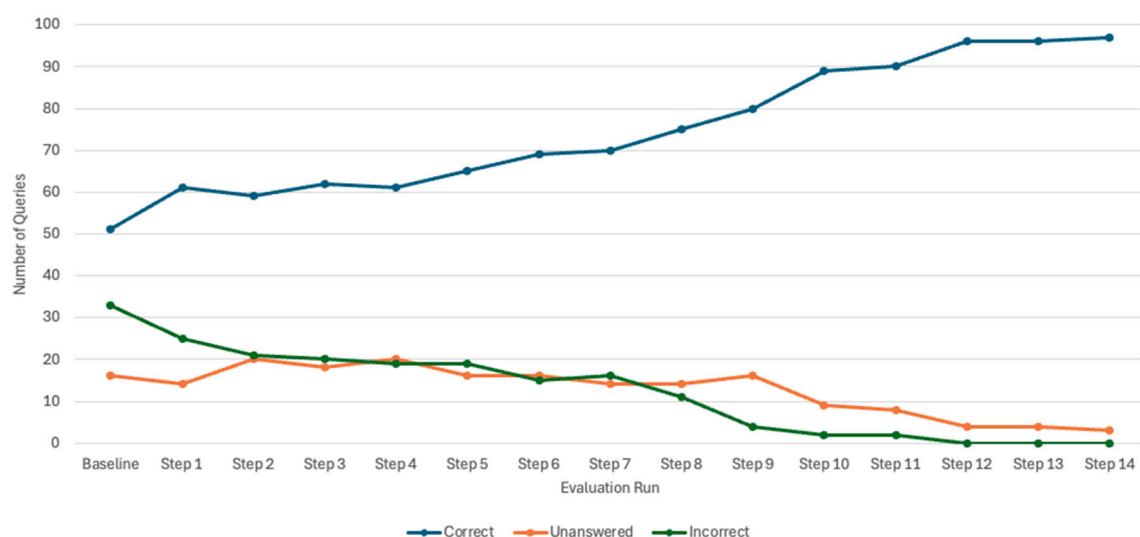
3.3. Changes Observed Across System-Level Refinements

The iterative evaluation showed that response reliability improved across multiple system-level refinements. The refinements addressed different components of the architecture, including structured domain information, semantic matching, entity and category processing, linguistic normalization, and response logic.

Three refinements were associated with notable changes in query outcomes. The provision of structured brand information to the generative model was associated with improved responses for 12 queries. Revising category-processing logic to prevent product names from being incorrectly interpreted as merchant categories was associated with seven queries' improvements. Improved handling of unrecognized geographical information was associated with changes in five queries.

Other refinements were accompanied by smaller improvements across the evaluation process. These included enriching lexical dictionaries, expanding example expressions for semantic matching, differentiating merchant subtypes, distinguishing unknown information from confirmed absence, improving spelling and linguistic normalization, and revising response logic. Overall, the successive refinements were accompanied by a progressive reduction in the number of incorrect and unanswered responses. These observations should not be interpreted as estimates of the independent causal contribution of each refinement because the refinements were introduced sequentially and cumulatively rather than evaluated through an ablation design.

Figure 2 presents the response outcomes across the baseline configuration and the 14 successive system refinements, illustrating how the response behavior evolved throughout the process.

**Figure 2.** Responses across the baseline configuration and 14 iterative system refinements using the fixed 100-query evaluation set.

3.4. Stability of Response Outcomes Across System Configurations

The evaluation records also showed differences in the stability of the response across the iterative refinement process. Queries that were ultimately resolved through the data-

grounded pathway exhibited fewer changes in response classification than those that remained in the language-model-generated pathway.

For the 90 queries handled through the data-grounded pathway in the final configuration, the mean number of changes in response outcome across the evaluation runs was 0.78 per query. In contrast, the remaining 10 queries in the language-model-generated pathway exhibited an average of 2.90 response outcome changes per query.

Among the 10 queries with the highest number of response outcome changes, eight were handled through the language-model-generated pathway, whereas two were resolved through the data-grounded pathway. Thus, queries remaining in the generative pathway showed greater variability in response classifications across successive system configurations. This analysis is descriptive and reflects stability across the observed system configurations; it does not constitute an ablation analysis of the individual refinements.

4. Discussion

4.1. System-Level Refinement and Response Reliability

The findings suggest that system performance is shaped not only by the capabilities of the underlying language models but also by the architecture through which domain knowledge is represented, interpreted, and transformed into responses. Previous studies on task-oriented dialogue have similarly emphasized the importance of structured dialogue states, domain knowledge, and external grounding for producing accurate and contextually appropriate responses [2,6,7,16,17]. For example, ontology-aware task-oriented dialogue systems explicitly represent domain-slot-value relationships to connect user constraints with domain databases, while knowledge-structured approaches have shown that incorporating relationships within external knowledge can improve dialogue-system inference.

This study extends this perspective by demonstrating that refinements to the processing mechanisms surrounding unchanged AI models can substantially improve response reliability. In particular, improvements emerged from better domain information representation, lexical and entity processing, category handling, reconciliation, and controlled response construction. This differs from approaches in which performance gains are primarily pursued through model fine-tuning or increased generative capability. Previous research has shown that supplying relevant contextual knowledge to language models can improve entity accuracy and reduce hallucination in task-oriented dialogue [6,7,12]. Simultaneously, evaluations of grounded LLM responses indicate that generative models may still introduce information that is not supported by the grounding material.

Therefore, these findings support a broader architectural interpretation of reliability. For structured service environments, the relevant design question is not only which model should be used but also which decisions should be delegated to the model and which should remain constrained by structured domain information and explicit system logic. This interpretation is consistent with recent modular approaches that combine LLM capabilities with symbolic planning and grounded execution rather than relying exclusively on end-to-end generation [7,18–20]. The iterative refinements indicate that improving the organization and use of domain knowledge can be as important as the generative capability of the model itself.

4.2. Selective Generative AI Use in Hybrid Architectures

The reduction in generative model dependency observed in this study supports a more selective role of generative AI in task-oriented systems. Existing research on LLM-based and task-oriented assistants has emphasized the importance of grounding, structured information, and controlled interaction with external resources to improve the reliability of generated responses [6,7,12]. Nevertheless, reliability and trustworthiness

remain important concerns when generative models are used in task-oriented settings [6]. These concerns are particularly relevant in financial applications, where the growing adoption of generative AI introduces opportunities alongside risks related to accuracy, trust, privacy, and responsible deployment [8–10,21].

The observed shift toward the data-grounded pathway suggests that the production of generative responses can be treated as a conditional architectural capability rather than a mandatory stage of the dialogue pipeline. This interpretation is consistent with modular and routing-based approaches that determine how LLM capabilities are invoked within larger systems [7,18–20].

This interpretation is reinforced by the differences in the stability of the pathway observed during the iterative evaluation. Requests ultimately handled through the data-grounded pathway showed fewer changes in response classification than those remaining dependent on generative response production. This observation is relevant to recent evidence that LLM outputs can exhibit non-deterministic behavior even under nominally deterministic system settings [22]. However, the present findings should not be interpreted as evidence that deterministic response generation is universally superior to LLM-based generation. In a structured digital service context, requests whose answers can be derived directly from institutional data may benefit from a more constrained response pathway, while generative models retain value for requests requiring greater linguistic flexibility or for cases that cannot be resolved through the available structured information.

4.3. Design Knowledge from the DSR Process

From a DSR perspective, the study's contribution extends beyond the performance of the final artifact. DSR emphasizes that artifact evaluation should generate knowledge about the design choices that contribute to its utility, rather than merely demonstrating that a technological solution works [13–15,23]. In this study, the iterative evaluation process provides such design knowledge by showing how modifications to different components of the hybrid architecture influenced the behavior of the assistant. The refinements did not contribute equally: changes related to structured domain information, category interpretation, entity handling, and linguistic normalization produced different effects across the evaluation process.

Three design principles can be derived from these observations. First, structured domain knowledge should be exploited before the generative response production is invoked. When the information required to resolve a request is already available within institutional data, the architecture should prioritize retrieving and organizing this information over delegating factual response construction directly to a generative model. Second, language understanding and response generation should be treated as separate functions. Semantic AI can remain responsible for interpreting diverse user expressions, while structured data and deterministic logic can support response construction whenever sufficient evidence is available. Third, in structured task-oriented environments, generative models should serve as a complementary fallback capability rather than the default response mechanism. This preserves their flexibility for unresolved requests while allowing more constrained mechanisms to handle requests for which the relevant information is already known. Additionally, iterative evaluations revealed that negative prompting alone was insufficient; reliable performance required either explicit domain data or bypassing the generative model via routing.

These principles also illustrate the value of iterative artifact evaluation [15,23]. The refinement process revealed that improving individual components in isolation was insufficient and that effective system performance depended on the interaction among semantic interpretation, structured knowledge, reconciliation, routing, and response con-

struction. Therefore, in DSR terms, the resulting contribution can be understood as both an evaluated artifact and a set of empirically derived design propositions for hybrid AI assistants [13–15,23,24]. However, because these propositions originate from a single digital financial service context, they should be regarded as context-grounded design principles requiring further evaluation, rather than universally applicable prescriptions.

4.4. Implications for Sustainable AI Assistants

The findings also have implications for the design of more sustainable AI assistants. Green AI research has increasingly emphasized that the environmental implications of AI should be considered alongside model performance and computational efficiency [3,4]. As generative AI systems become more widely deployed, attention has increasingly extended to the resource requirements associated with their operational use and to strategies for improving the efficiency of generative AI systems [5,25–27]. Recent work on green prompt engineering further demonstrates that the way generative models are invoked and prompted can influence their resource requirements [5]. These developments indicate that sustainable AI deployment requires attention to model efficiency and how generative capabilities are used in operational systems. The relationship between generative AI and sustainability is increasingly being examined across diverse application domains, including sustainable and inclusive education and AI-supported green consumption [28,29].

From this perspective, the contribution of the proposed architecture lies in the selective utilization of the generative model rather than the demonstrated computational or environmental efficiency. In the final system configuration, 90% of the evaluated queries were resolved without using Qwen3:8B to generate the final response, whereas 10% remained dependent on generative response production. This does not imply that these interactions were computationally cost-free. The 566-million-parameter bge-m3 semantic model remained part of the processing pipeline, and additional operations, including entity and parameter processing, reconciliation, routing, structured-data retrieval, and deterministic response construction, also required computational resources.

Consequently, the observed reduction represents a shift in generative model dependency rather than a direct measurement of computational-resource savings or environmental optimization. This study did not directly measure electricity consumption, GPU utilization, inference latency, token-level computational cost, operational cost, or carbon emissions. Accordingly, whether reduced generative-model use translates into measurable environmental or operational benefits requires a direct empirical evaluation. Therefore, the present findings position hybrid routing as a potential resource-conscious architectural strategy rather than a demonstrated environmental optimization.

LLM-based approaches are already being investigated in sustainability-oriented applications and decision-support settings [24,26]. However, the present study addresses sustainability from a different perspective by examining whether generative capabilities need to be invoked for every request. Therefore, future evaluations should combine the response-level measures used in this study with direct measurements of energy consumption, latency, generated tokens, computational utilization, and, where possible, associated carbon emissions. Such evaluation would make it possible to determine whether reduced generative-model dependency translates into measurable environmental and operational benefits under real deployment conditions.

4.5. Limitations and Future Research

Several limitations should be considered when interpreting the study findings. First, the evaluation was conducted using a fixed set of 100 queries within a specific digital payment environment. Although keeping the query set unchanged was important for

comparing successive refinements, the resulting performance should not be assumed to generalize directly to other financial services, domains, user populations, or conversational settings. The fixed-query design supports controlled within-system comparison across successive configurations but does not allow for establishing out-of-sample generalizability. Moreover, because the same query set was repeatedly used across the 14 refinement cycles, later system modifications may have become increasingly adapted to this evaluation set, creating a risk of overfitting and optimistic performance estimates. This consideration is particularly relevant in DSR, where artifact evaluation is closely tied to the context in which the artifact is developed and tested. Further evaluation is important for establishing broader applicability.

Second, the study evaluated a specific hybrid architecture using bge-m3 for semantic matching and Qwen3:8B for generative processing. The underlying models remained unchanged during the refinement process. Consequently, the present evaluation cannot fully separate architecture-level effects from effects that may be specific to this particular model combination. Moreover, the study does not establish whether the same patterns would emerge with alternative embedding models, generative models, or different model sizes. Therefore, future research should replicate the architecture with different model combinations and extend the evaluation to larger, independently constructed, and out-of-sample query sets, including more diverse linguistic expressions and interaction scenarios.

Third, the evaluation focused on factual response correctness and response-pathway behavior rather than subjective aspects of user experience. The fluency, naturalness, perceived helpfulness, and user satisfaction associated with template-based deterministic responses were not compared with those of language-model-generated responses. In addition, the 14 refinements were introduced sequentially and cumulatively rather than through a controlled ablation design. Consequently, the evaluation does not isolate the independent causal contribution of each refinement. Therefore, future studies should complement factual correctness assessment with user-centered evaluation and controlled ablation experiments to examine both perceived response quality and the individual contribution of architectural refinements.

5. Conclusions

This study investigated whether a hybrid, data-grounded architecture could improve the reliability of a task-oriented AI assistant while reducing its dependence on generative response production. The artifact was evaluated using a fixed set of 100 realistic queries and progressively refined through fourteen system-level modifications. Correct responses increased from 51% to 97%, whereas incorrect responses decreased from 33% to zero. Concurrently, queries requiring generative response production dropped sharply from 34% to 10%, achieving a 70.5% relative reduction in generative model invocations. These improvements were achieved without replacing the underlying semantic and generative models, highlighting the role of system-level mechanisms, such as structured domain knowledge, entity and category processing, reconciliation, routing, and controlled response construction. These findings address the three research questions. Regarding RQ1, the iterative system-level refinements substantially improved the correctness of the response. Regarding RQ2, the hybrid architecture shifted response production toward the data-grounded deterministic pathway and reduced the reliance on generative responses. Regarding RQ3, queries remaining in the generative pathway exhibited greater response-outcome variability across successive system configurations than those ultimately resolved through the data-grounded pathway.

The findings indicate that improving task-oriented AI assistants does not necessarily require greater reliance on generative response production. In structured digital financial

services, generative models can be used selectively as a fallback, while requests supported by sufficient domain information can be resolved through deterministic, data-grounded mechanisms. This hybrid approach provides a practical direction for combining response reliability and flexibility with more resource-conscious use of generative AI.

Author Contributions: Conceptualization, B.O.K., V.C. and U.B.; methodology, B.O.K. and V.C.; software, U.B.; validation, B.O.K., V.C. and U.B.; formal analysis, B.O.K. and U.B.; investigation, B.O.K., V.C. and U.B.; resources, U.B.; data curation, U.B.; writing—original draft preparation, B.O.K. and V.C.; writing—review and editing, B.O.K. and V.C.; visualization, B.O.K. and U.B.; supervision, B.O.K. and V.C.; project administration, V.C. and U.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to the confidentiality requirements and approval of the industrial partner. The data are not publicly available due to confidentiality and proprietary restrictions associated with the evaluated commercial system.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.6) for language refinement, including grammar, clarity, and readability. The authors reviewed and edited the output and take full responsibility for the content of the publication.

Conflicts of Interest: Author U.B. is an employee of Konfides Information Technologies, the company associated with the development of the system examined in this study. The other authors declare no conflicts of interest. The authors retained full responsibility for the analysis and interpretation of the results and for the preparation of the manuscript.

References

1. Dong, X.L.; Moon, S.; Xu, Y.E.; Malik, K.; Yu, Z. Towards next-generation intelligent assistants leveraging llm techniques. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; pp. 5792–5793.
2. Sánchez Cuadrado, J.; Pérez-Soler, S.; Guerra, E.; De Lara, J. Automating the Development of Task-oriented LLM-based Chatbots. In Proceedings of the 6th ACM Conference on Conversational User Interfaces, Luxembourg, 8–10 July 2024; pp. 1–10.
3. Verdecchia, R.; Sallou, J.; Cruz, L. A systematic review of Green AI. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2023**, *13*, e1507. [[CrossRef](#)]
4. Bolón-Canedo, V.; Morán-Fernández, L.; Cancela, B.; Alonso-Betanzos, A. A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing* **2024**, *599*, 128096. [[CrossRef](#)]
5. Podder, S.; Date, H.; Murthy, S. Green prompt engineering for sustainable generative ai. *Environ. Sci. Ecotechnol.* **2026**, *30*, 100684. [[CrossRef](#)] [[PubMed](#)]
6. Labruna, T.; Brenna, S.; Bonetta, G.; Magnini, B. Are you a good assistant? Assessing llm trustability in task-oriented dialogues. In Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, 4–6 December 2024; pp. 470–477.
7. Xu, H.D.; Mao, X.L.; Sun, F.; Che, T.Y.; Xu, C.; Huang, H. Agenttod: A task-oriented dialogue agent with a flexible and adaptive api calling paradigm. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–32. [[CrossRef](#)] [[PubMed](#)]
8. Dahiphale, D.; Madiraju, N.; Lin, J.; Karve, R.; Agrawal, M.; Modwal, A.; Balakrishnan, R.; Shah, S.; Kaushal, G.; Mandawat, P.; et al. Enhancing trust and safety in digital payments: An LLM-powered approach. In *2024 IEEE International Conference on Big Data (BigData)*; IEEE: New York, NY, USA, 2024; pp. 4854–4863.
9. Easin, A.M.; Sourav, S.; Tamás, O. An intelligent llm-powered personalized assistant for digital banking using langgraph and chain of thoughts. In *2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY)*; IEEE: Pula, Croatia, 2024; pp. 625–630.
10. Henni, M.D. Large Language Models in Digital Banking: A Systematic Review and Implications for Financial Inclusion (2015–2025). *IJEP* **2026**, *9*, 20–52.

11. Gudimella, J.; Chittari, R.V.; Munigadapa, L.S.; Tadepalli, S.K. Finance Utility LLM: A Multi-Agent LLM Framework for Intelligent Personal Finance Management. In *International Conference on Edge Computing and Applications*; Springer Nature: Cham, Switzerland, 2025; pp. 392–404.
12. Elnashar, A.; White, J.; Schmidt, D.C. Prompt engineering for structured data: A comparative evaluation of styles and LLM performance. *Artif. Intell. Auton. Syst.* **2025**, *2*, 32–49. [[CrossRef](#)]
13. Hevner, A.; Chatterjee, S. Design science research in information systems. In *Design Research in Information Systems: Theory and Practice*; Springer: Boston, MA, USA, 2010; pp. 9–22.
14. Vom Brocke, J.; Hevner, A.; Maedche, A. Introduction to design science research. In *Design Science Research. Cases*; Springer International Publishing: Cham, Switzerland, 2020; pp. 1–13.
15. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. A design science research methodology for information systems research. *J. Manag. Inf. Syst.* **2007**, *24*, 45–77. [[CrossRef](#)]
16. Wang, Z.; Geng, N.; Guo, Z.; Ma, W.; Zhang, M. Human vs. Agent in Task-Oriented Conversations. In Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, Xi'an, China, 7–10 December 2025; pp. 133–142.
17. Dong, A.; Li, P.; Chen, Y.; Gibson, S.; Zhao, L.; He, M. Human–AI Collaboration Across Decision Support, Autonomous Systems, and LLM Agents: A Systematic Review and Collaboration Convergence Framework. *Sustainability* **2026**, *18*, 5313. [[CrossRef](#)]
18. Manias, D.M.; Chouman, A.; Shami, A. Semantic routing for enhanced performance of llm-assisted intent-based 5g core network management and orchestration. In *GLOBECOM 2024-2024 IEEE Global Communications Conference*; IEEE: Cape Town, South Africa, 2024; pp. 2924–2929.
19. D'Urso, S.; Martini, B.; Sciarrone, F. A novel LLM architecture for intelligent system configuration. In *2024 28th International Conference Information Visualisation (IV)*; IEEE: Coimbra, Portugal, 2024; pp. 326–331.
20. Schnieder, M.; Schimanke, F.; Sert, M.; Persia, F.; Lyu, C.; Glesene, G.B. Semantic Routing: Fusing LLMs and Combinatorial Optimization Problems within the Generative Problem Solving (GPS) Domain. *Int. J. Semant. Comput.* **2026**, *20*, 351–370. [[CrossRef](#)]
21. Mahmud, M.A.; ALAM, M.A.; Alam, M.K. Exploring the transformative potential of generative ai and large language models (llms) in financial applications: Opportunities, risks and strategic implications. *J. Comput. Sci. Technol. Stud.* **2025**, *7*, 1069–1088.
22. Atıl, B.; Aykent, S.; Chittams, A.; Fu, L.; Passonneau, R.J.; Radcliffe, E.; Rajagopal, G.R.; Sloan, A.; Tudrej, T.; Ture, F.; et al. Non-determinism of “deterministic” LLM system settings in hosted environments. In Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems, Mumbai, India, 23 December 2025; pp. 135–148.
23. Hevner, A.R.; Parsons, J.; Brendel, A.B.; Lukyanenko, R.; Tiefenbeck, V.; Tremblay, M.C.; vom Brocke, J. Transparency in design science research. *Decis. Support Syst.* **2024**, *182*, 114236. [[CrossRef](#)]
24. Giudici, M.; Abbo, G.A.; Belotti, O.; Braccini, A.; Dubini, F.; Izzo, R.A.; Crovari, P.; Garzotto, F. Assessing llms responses in the field of domestic sustainability: An exploratory study. In *2023 Third International Conference on Digital Data Processing (DDP)*; IEEE: Luton, UK, 2023; pp. 42–48.
25. Goel, P.K.; Yadav, S.P.; Upadhyay, P. Sustainability in Multi-Agent LLM System: Energy Efficiency and Green AI Initiatives. In *Advancements in Multi-Agent Large Language Model Systems for Next-Generation AI*; IGI Global Scientific Publishing: Hershey, PA, USA, 2026; pp. 189–208.
26. Arslan, M.; Munawar, S.; Sibilla, M. Sustainable energy decision-making with an RAG-LLM system. In *2024 International Conference on Decision Aid Sciences and Applications (DASA)*; IEEE: Manama, Bahrain, 2024; pp. 1–5.
27. Khan, H.R.; Raj, R.R.; Khan, A.R.; Badrolhisam, N.I. Leveraging Generative AI for Sustainable Development: Opportunities, Risks and Ethical Pathways. *J. Bus. Soc. Sci.* **2025**, *2025*. [[CrossRef](#)]
28. Adauto-Medina, W.; Chamorro-Atalaya, O.; Olivares-Zegarra, S.; Arévalo-Tuesta, J.A.; Arones, M.; Aybar-Bellido, I.; León-Velarde, C.; Fernández-Flores, S.; Quispe-Andía, A.; Auqui-Ramos, E. From Multimodal Texts to Generative AI: A Systematic Review of Immersive Educational Strategies and Their Reported Contributions to Sustainability and Inclusion in Higher Education. *Sustainability* **2026**, *18*, 6373. [[CrossRef](#)]
29. Si, K.; Wang, C.; Wei, S.; Lan, Y. Generative AI-Enabled Precision Recommendation for Green Products: Mechanisms of Consumer Cognitive Fluency and Low-Carbon Purchase Decisions. *Sustainability* **2026**, *18*, 2018. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.