Cloud Computing and Data Science
https://ojs.wiserpub.com/index.php/CCDS/

UNIVERSAL WISER PUBLISHER

Research Article

The Personalization Paradox: Semantic Loss Vs. Reasoning Gains in Agentic AI Q & A

Satyajit Movidi[1*], **Stephen Russell**[2]

[1]Department of Mathematics and Statistics, University of West Florida, Pensacola, FL, 32514, USA
[2]Department of Intelligent Systems and Robotics, University of West Florida, Pensacola, FL, 32514, USA
E-mail: sm402@students.uwf.edu

Received: 8 January 2026; **Revised:** 20 April 2026; **Accepted:** 12 May 2026

Abstract: This study examines how personalization in agentic retrieval-augmented Artificial Intelligence (AI) systems influences the quality of answers delivered in institutional knowledge access settings such as academic advising. Prior advising and knowledge-access systems typically assume personalization is universally beneficial, yet little empirical evidence evaluates how it alters information quality. This paper addresses this gap by analyzing personalization as an independent factor within a Retrieval-Augmented Generation Large Language Model (RAG LLM) used for student advising. The study evaluates ten system configurations across personalized and non-personalized conditions using twelve authentic advising questions intentionally designed for lexical strictness. Performance was assessed using lexical metrics (Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-L), semantic similarity measures (Metric for Evaluation of Translation with Explicit ORdering (METEOR), BERTScore), and reasoning/grounding metrics from the Retrieval Augmented Generation Assessment (RAGAs) framework. A Linear Mixed-effects Model (LMM) was used to quantify main effects and interactions. Personalization in agentic AI does not yield uniform gains; instead, it creates a critical trade-off where factors that significantly improve reasoning quality also incur a statistically significant penalty on semantic similarity. Specifically, personalized configurations produced a statistically significant decrease in BERTScore (0.841 vs. 0.848, $p < 0.0001$) alongside a simultaneous and significant improvement in METEOR (0.361 vs. 0.251, $\Delta = +0.110$, $p < 0.0001$), together demonstrating the metric-dependent nature of the trade-off. Grounding and reasoning metrics simultaneously improved, with Faithfulness rising from 0.655 to 0.711 ($p = 0.0135$), further supporting that personalization enhances answer quality even as it is penalized by semantic similarity metrics. Personalization decreases semantic similarity scores, not due to quality loss but because generic semantic metrics penalize beneficial user-specific deviations. The configuration that applied personalization redundantly across all three stages: setting the AI's role, guiding document retrieval, and conditioning the final response generation, achieved the best overall results, confirming that fully integrated user-specific adaptation yields the most effective balance between reasoning gains and semantic penalties.

Keywords: personalization, knowledge access systems, retrieval augmented generation, large language models, semantic similarity, agentic Artificial Intelligence (AI), academic advising

DOI: https://doi.org/10.37256/ccds.7220269566

1. Introduction

The adoption of large language models and agentic Artificial Intelligence (AI) systems is accelerating in organizations that depend on complex Knowledge Bases (KB), including higher education institutions. Student-facing services such as academic advising increasingly rely on institutional policies, catalogs, and frequently asked questions as primary knowledge sources. Students expect on-demand guidance, yet traditional advising remains uneven in availability and places pressure on faculty and staff to meet student needs and expectations. AI-enabled advising and knowledge access tools promise to expand access and reduce advisor burden, a critical need given increasing advising workloads and institutional pressures [1]. Many systems today rely on generic Retrieval Augmented Generation (RAG) pipelines that provide factually correct but context-agnostic responses. Such systems rarely integrate personalization in the form of a student's academic history, goals, or constraints, creating risks of disengagement or misalignment with individual needs [2]. Personalization can provide a mechanism that conditions responses based on users' intent, role, or context so that system-provided answers are both correct and tailored to the user.

Within the broader information and knowledge management literature, personalization is often viewed as a way to filter, prioritize, or tailor access to organizational knowledge resources. While research on personalization in AI has expanded substantially in fields such as recommendation systems and affective computing [3, 4], higher education advising and similar institutional knowledge access scenarios have received considerably less attention within that scope. The challenge of aligning answers with student intent, situated background, or personal context remains largely unexamined in advising-oriented knowledge access systems. Academic advising, however, presents a uniquely high-stakes instantiation of this challenge: unlike a product recommendation or a marketing message, an incorrectly personalized advising response can directly affect a student's academic trajectory, credit completion, or graduation timeline. This makes it not only a natural next domain for personalization research but arguably one of the most consequential.

Intuitively, many stakeholders believe personalization is always beneficial and should be built into knowledge access tools by default. It is also becoming increasingly integrated into modern AI systems. Yet prior studies on AI advising or question answering have largely emphasized descriptive outcomes, leaving open the theoretical and methodological question of when personalization is appropriate and how robust its effects are under different evaluation regimes. To investigate the implications of user-specific adaptation on student advising as an AI knowledge access problem, we developed AiVisor, a prototype agentic AI advising system that operates over institutional policies and frequently asked questions and was developed. The study's hypothesis is that personalization matters because it shifts the alignment and reasoning profile of answers, and AiVisor is used to test the nature and direction of this shift across different metric families. Effects are measured across similarity and semantic metrics (Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-L, Metric for Evaluation of Translation with Explicit ORdering (METEOR), BERTScore) and also through rigorous statistical analysis to assess the significance of observed differences. Importantly, AiVisor is evaluated on a conservative test setting. The question set is intentionally skewed toward lexical and generic phrasing, emphasizing exact wording over open-ended synthesis. This constitutes a worst-case evaluation, designed to reveal the trade-offs that arise when user-tailored responses deviate from a single, generic ground truth.

AiVisor distinguishes itself from prior advising and question answering research by making personalization the central experimental variable rather than an assumed benefit. The study shows that the effects of personalization vary by question type, with gains in some cases and negligible or even negative impacts in others. Unlike earlier work that relied on small pilots or perception surveys, AiVisor employs multi-metric, semantic, and statistical evaluation, bridging technical system analysis with user-centered knowledge access outcomes. For the information and knowledge management community, the study contributes both an empirical characterization of personalization effects in an institutional setting and a methodological template for multi-metric evaluation of retrieval-augmented systems.

The remainder of the paper proceeds as follows. The next section reviews related work on personalization, knowledge access, and AI student advising systems. This is followed by a description of the AiVisor system design, details of the experimental methodology, and a presentation of the results. The remaining sections discuss implications for knowledge management and system design; limitations and future work; and conclusions.

2. Background and related work

Personalization has become a cornerstone of modern AI, enabling systems to tailor interactions to individual users' preferences, context, and goals [5]. Large Language Models (LLMs) in their vanilla form excel at general knowledge tasks but struggle with user-specific adaptation. For example, they have difficulty understanding an individual's emotions, style, or preferences without additional context. This gap has spurred research into personalized LLMs and RAG systems that leverage user data (profiles, history, etc.) to produce responses more relevant to each user [6]. Across domains like marketing and human-computer interaction, personalization is recognized as key to user engagement. In marketing, for instance, it is considered a key strategy for delivering relevant customer experiences, though it requires deep understanding of user needs and behaviors [7]. Similarly, AI-driven services such as chatbots and recommender systems increasingly incorporate context-aware guidance to improve user satisfaction and effectiveness. A survey by [5] for example, systematically examines how individualized generation can be integrated at all stages of RAG, yet it also emphasizes that prior work has not comprehensively addressed personalized RAG. Most existing surveys have focused on general RAG or agents without systematically exploring their implications for user-tailored performance. Prior work in human-AI interaction has further emphasized that personalization plays a meaningful role in producing more contextually and semantically appropriate responses, with user-specific adaptation improving both relevance and communicative fit [8, 9]. In short, while personalization is widely pursued in AI, its integration with advanced frameworks like RAG is still an evolving research area.

In adjacent domains, such as chatbots and recommender systems, personalization has been studied for years. For example, Blömker and Albrecht [10] examine how few-shot methods can craft distinct conversational personas in service chatbots, enhancing user engagement and perceived alignment with the user. The notion of few-shot personalization of LLMs is also reflected in recent work: Kim and Yang [11] propose a system that iteratively adjusts prompts per user using "mis-aligned responses" to improve user-tailored performance.

Table 1. Brief summary of adjacent domain literature

Reference	Domain/Context	Methods/Approach
Li et al. [5]	LLM/RAG personalization	Surveys personalization techniques across pre-retrieval, retrieval and generation stages of RAG.
Akiba and Fraboni [12]	Academic advising/ChatGPT	Empirical study of ChatGPT on academic advising queries across a range of student scenarios.
Antico et al. [13]	RAG-based advising chatbot	Built and piloted a RAG-driven chatbot over institutional documents for academic advising.
Blömker and Albrecht [10]	Service/Customer chatbots	Applies few-shot learning to craft distinct conversational personas in service chatbots.
Kim and Yang [11]	LLM personalization	Proposes Fermi, a few-shot prompt tuning system that iteratively refines personalization using misaligned responses.
Siddique et al. [14]	Dialog systems personalization	Introduces a zero-shot reward-based framework for personalizing task-oriented dialogue without per-user labeled data.
Chandra et al. [7]	Marketin/Digital personalization	Reviews trends in personalized marketing including recommender systems, segmentation, and dynamic content adaptation.
Casaca and Miguel [15]	Consumer research	Systematically reviews personalization's impact on consumer satisfaction, examining moderating factors such as privacy concerns.
Tarifi and Bakhsh [16]	Marketing strategy/Applied	Empirically and strategically examines personalization tactics including recommendation engines and dynamic content strategies.

Similarly, personalization is not just a superficial feature (e.g. calling the user by their name), but involves modeling preferences, past interactions, and adapting system behavior over time. This research focuses on user-specific adaptation can be seen in the dialog system proposed by [14], where they illustrate a zero-shot reward-based personalization framework for task-oriented dialogue systems, enabling adaptation without per-user labeled data. Recent literature reviews also document trends in personalized AI marketing, identifying major themes like recommendation engines, customer segmentation, and real-time individualization [7, 15, 16]. Table 1 provides a summary of work in various domains.

2.1 *Personalization and question & answer systems*

Contemporary personalization studies often evaluate on prompts that obviously benefit from context or they measure with a single metric, risking overfitting to that measure. Few papers normalize across metrics, report paired statistics, or ask whether gains persist when questions are generic and strictly lexical. AiVisor addresses these gaps by using a lexically strict evaluation suite, multi-metric z-score normalization, and paired tests with effect sizes. This combination is designed to separate true semantic improvements from superficial paraphrase and clarifies where lexical penalties appear and how large they are.

Most Question & Answer (Q & A) benchmarks and leaderboards optimize for generic performance. Datasets such as Stanford Question Answering Dataset (SQuAD) [17], Microsoft Machine Reading Comprehension (MS MARCO) [18], and Natural Questions [19] largely omit user-specific context. While recent LLM-based systems show promise [20], without grounding, they risk hallucination and a lack of specificity. RAG approaches [21] combine retrieval with LLMs to improve factuality, and subsequent work has demonstrated that these techniques meaningfully improve factual accuracy and knowledge grounding across diverse retrieval settings [22]. However, the literature still offers limited, rigorous evidence for user-centric personalization in Q & A systems, especially under lexical stringency. This gap is often measured using standard evaluation metrics like BLEU [23], ROUGE-L [24], METEOR [25], and BERTScore [26], which provide complementary views of lexical and semantic overlap.

2.2 *Student advising systems*

Thottoli et al. [27] conducted a literature review and found only 67 relevant papers, from several hundred search results, covering the period from 1984-2023 on chatbots or AI in academic advising. For example, Lin and Yu [28] performed a bibliometric analysis of educational chatbots, and Bilquise et al. [29] proposed a bilingual advising chatbot. Their review highlighted that this research was spread across several domains (e.g., education technology, computer science, management, and information systems) and disparate topics (e.g., student engagement, Frequently Asked Question (FAQ) answering, enrollment support, or counseling), lacking a cohesive research agenda or connections to build cumulative knowledge. They also found a lack of common or unified frameworks, and shallow empirical validation. They concluded that a minimal amount of study has been done on applying chatbots and AI for academic advising in higher education institutions, identifying this as a major research gap.

Dawood [30] provides a narrative literature review on the use of chatbots for personalized academic advising in higher education, synthesizing findings from 20 recent studies (2020-2024). This work defines personalization as tailored responses beyond generic frequently asked questions, recommender-based suggestions, and dynamic adaptation in multi-turn dialog. Dawood's review [30] concluded that while "personalized" chatbots hold transformative potential for scaling advising services, especially in reducing advisor workloads, current systems remain limited in handling complex advising. However, Dawood [30] emphasizes that most existing systems claim context-aware guidance capabilities, but in practice, many fall short, delivering only generic or rule-based answers. The review highlights this as one of the field's biggest gaps.

Despite the sparsity of work found by [27], and the limitations found by [30], there is some literature directly related to RAG-based LLM student advising systems. For example, Akiba and Fraboni [12] examined ChatGPT's potential in academic advising and remarked that this context does not appear to have been very well studied so far. They point out that much of the early discourse around tools like ChatGPT in higher education has focused on cheating or broad pedagogical impact, while student support services (like advising) were comparatively neglected. Antico et al. [13] constructed a RAG-driven chatbot system by ingesting institutional documents and customizing GPT prompts for

the University of Milano-Bicocca. They ran a small usability pilot ($n = 6$) and found positive responses to the interface and conversational tone. However, they also observed that the system sometimes failed to provide fully accurate answers, omitted relevant document content, and generated un-clickable or broken link issues. Another work by [31] introduced Advisely, a GPT-4-powered academic advising platform designed to address the common challenges of traditional advising such as inconsistent guidance, limited accessibility, and high advisor workloads. Abdelhamid et al. [31] demonstrated that institution-specific knowledge integration combined with LLM capabilities can improve advising accessibility, accuracy, and consistency.

Soomro et al. [32] examined how AI-driven academic advising systems can support student success by improving retention, enhancing academic confidence, and aligning educational choices with career pathways. Their study surveyed 1,200 university students across three institutions that had deployed AI-powered advising platforms. The AI advising systems studied by [32] combined Natural Language Processing (NLP), machine learning-based recommendation engines, predictive analytics, and conversational chatbots integrated with institutional data, enabling tailored course advice and career-aligned guidance. Unlike architecture-specific implementations (e.g., agentic RAG or few-shot personalization), their focus was on the functional composition of advising platforms and students perceived personalization benefits. This survey-based evaluation of perceived outcomes contrasts with the present study, which operationalizes personalization via algorithmic adaptation (agentic RAG) and evaluates its effects using objective performance metrics.

Luong and Luong [33] present FIT-Advisor, a chatbot-based academic advising system developed for Information Technology (IT) students. Their approach integrates natural language processing, machine learning, and recommender algorithms to assist students with training regulations, curriculum information, basic IT knowledge, and recommended course suggestions. This study demonstrated how localized AI-driven advising systems can improve accessibility and accuracy in higher education. Luong and Luong [33] highlight that FIT-Advisor should complement rather than replace human advisors, reserving complex advising cases for human expertise while offloading routine advising tasks to the chatbot.

Another work by [34] is quite comprehensive in its contributions. This work extends an existing system based on Gemini 1 Pro and RAG over a university student handbook with an agent-based system built on GPT-4 with the text-embedding-ada-002 model. They integrated three tools: RAG (student handbook), Google Search (for out-of-scope queries), and Gmail (for human-in-the-loop escalation). Evaluation used 18 questions derived from the student handbook, scored with the Retrieval Augmented Generation Assessment (RAGAs) framework (answer correctness, context precision, faithfulness, entity recall, and answer relevancy). Their results showed that the agentic version outperformed the retrieval-only version across all metrics except entity recall, with strong performance in conversational flow, contextual relevance, and reliability. Moreover, this work incorporated human-in-the-loop and hallucination mitigation capabilities. However, this effort did not address personalization or lexical and semantic concerns.

These individual efforts show that the idea of AI advisors is emerging, but the integration of personalization into such systems remains largely unexplored. The literature review highlights that even widely cited papers in this area cover diverse topics with no clear demarcations. In short, user-tailored AI advising is an under-researched area, with only nascent studies pointing to its promise. This is an important observation for academics and practitioners: unlike domains such as marketing or e-commerce where user-specific techniques are relatively mature, in student advising the field is only beginning to ask fundamental questions and run early experiments. While prior research has explored AI based advising systems, most efforts have been either architectural demonstrations such as retrieval augmented generation versus agentic designs, small scale institutional pilots such as Advisely and FIT Advisor, or survey-based evaluations of perceived personalization. These studies highlight the potential of AI to scale advising, yet remain fragmented, often limited by small evaluation sets, self-reported outcomes, or narrow technical scope. AiVisor differentiates itself from prior work in four key ways:

(1) Personalization efficacy as the central research variable. Prior work has generally treated personalization as an assumed benefit or measured it indirectly through student surveys. AiVisor instead isolates personalization as the experimental factor, directly testing whether it improves or weakens advising responses.

(2) Question type sensitivity. AiVisor demonstrates that personalization does not uniformly improve outcomes. Some advising question types show strong gains, while others see little to no benefit or even negative impacts in others. This nuanced result challenges the assumption that personalization is universally positive and brings new theoretical

insight to advising research.

(3) Methodological rigor. Whereas previous studies often relied on small scale descriptive metrics or subjective feedback, AiVisor applies a suite of lexical and semantic similarity measures including BLEU, ROUGE-L, BERTScore, and METEOR, combined with statistical testing. This provides a more robust and quantitative evaluation framework than has been used in advising research to date.

(4) Bridging technical and user-centered evaluation. Prior studies tend to focus either on technical architecture comparisons or on user perceptions of usefulness. AiVisor positions itself in between, offering system-level evidence of personalization effects that can inform both design choices and user-facing outcomes.

Together, these contributions establish AiVisor as the first advising study to rigorously and systematically evaluate the real impact of personalization, providing both methodological innovation and theoretical differentiation from prior literature. Building on the gaps identified in this review, namely the absence of rigorous, multi-metric evaluation of personalization effects in advising systems, this study tests two hypotheses. H1 posits that personalization's effects are complex and metric-dependent, creating trade-offs between semantic, lexical, and reasoning quality rather than uniform improvement. H2 posits that personalization does not substantially alter lexical composition, indicating that any observed improvements occur at the level of meaning rather than expression. Before providing details on the experimental methodology, an overview of the system design is presented in the next section.

3. System design

The AiVisor system is designed around Gemini-Flash 1.5 LLM, an agent that retrieves personalization information, and a Facebook AI Similarity Search (FAISS) vector Database (vectorDB) RAG pipeline. Figure 1 presents a block illustration of AiVisor. Institutional documents (policies, catalogs, and FAQs) form the core retrieval corpus, while the user-personalization conditioning agent optionally injects user-specific content into either vectorDB or prompt assembly. Prompt assembly composes the LLM input, which includes a role prompt, the RAG context, user profile information, and the user's question. Some elements of the assembled prompt are included or excluded depending on the system configuration.

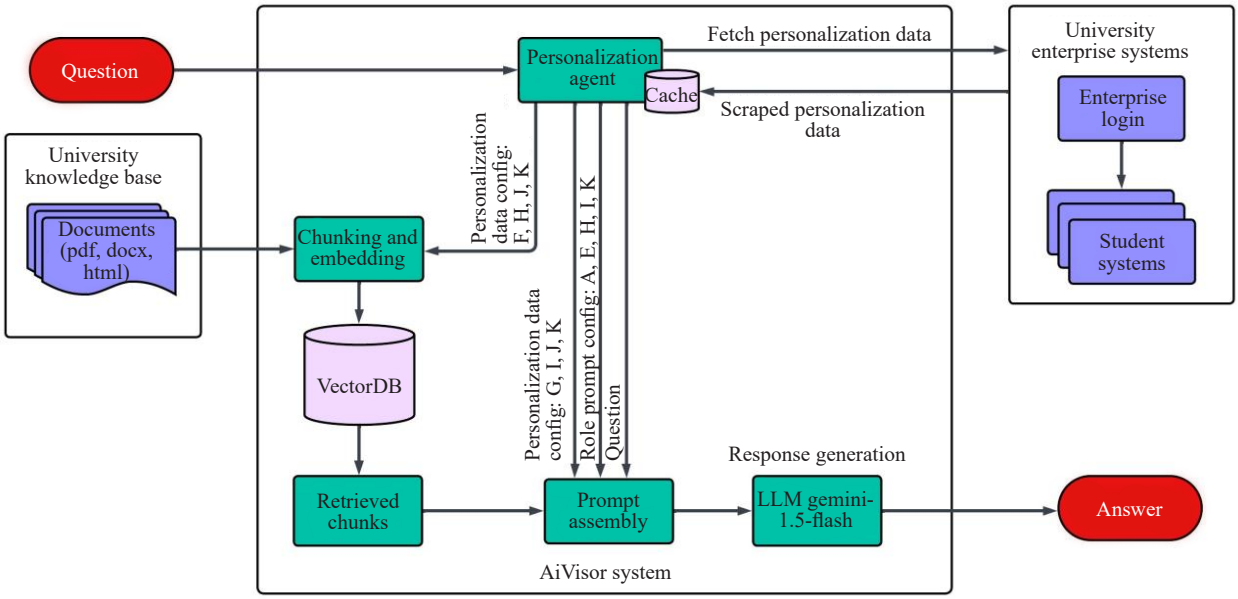


Figure 1. Block illustration of the AiVisor system with Personalization Agent, VectorDB, and Prompt Assembly

3.1 *Chunking configuration*

Documents were segmented into 10,000-character chunks with a 1,000-character overlap using a recursive text splitter. Gemini Flash 1.5 supports context windows over millions of tokens [35], enabling the use of larger retrieval units that preserve cross-section dependencies such as definitions, tables, and cross-references. Maintaining such coherence reduces hallucinations arising from fragmenting logically connected text [36]. Keeping quoted or cited material and its supporting context within a single retrieval unit also lowers the risk of accurate quotations being divorced from their qualifying details. In contrast, small chunks typically require large overlaps to prevent context breaks, inflating storage with near-duplicate vectors. Larger chunks, therefore, minimize both redundancy and fragmentation.

For retrieval, we used the FAISS vectorDB with the default $k = 4$, including all retrieved chunks in the prompt assembly. While coarse chunking improves the internal coherence of policy and FAQ passages, it can also blur topical boundaries, introduce irrelevant context, and reduce retriever precision. Consequently, the larger chunk size increases background noise and may partially obscure personalization signals. Because retrieval optimization was not the central focus of this study, the chosen coarse-grained chunking (10,000-character chunks) should be viewed as a conservative methodological condition that likely attenuates any observed benefit of personalization by fragmenting contextual relevance signals.

3.2 *Experimental system configurations*

Table 2. System configuration summary

Conf	Configuration	Description
A	Web baseline (No RAG)	Uses web results with a role prompt; serves as the non-RAG reference system.
B	Web + Personalization	Adds a personalization prompt to the web baseline for user-context adaptation.
C	Web-only (No role/Personalization)	Generates responses directly from the user query without role or personalization context.
D	VectorDB-only RAG baseline	Retrieves from internal sources (FAQ, policy, KB) with no prompts or personalization.
E	RAG with role prompt	Adds a role prompt to the RAG baseline for contextual, role-aware retrieval.
F	Embedded personalization in retriever	Integrates personalization within VectorDB retrieval, tailoring results at the retriever level.
G	RAG with external personalization prompt	Applies personalization at the generation stage, combining retrieved data with a personalization prompt.
H	Personalized retrieval with role context	Combines role input and embedded personalization in retrieval for contextualized and user-aware generation.
I	Dual-layer prompting (Role + Personalization)	Includes both role and personalization prompts alongside retrieved data for a multi-layer RAG approach.
J	Redundant personalization (Retriever + Prompt)	Personalization applied both within retrieval and generation prompts, resulting in overlapping user-context signals.
K	Full dual-layer personalization system	Combines role input, retriever-level personalization, and redundant prompt-level personalization; the most comprehensive configuration.

Ten system configurations (A-K) that varied in their use of RAG, role conditioning, and personalization were tested. Configuration B was omitted from testing to avoid any privacy concerns. Table 2 provides the details of each configuration. Configuration A served as a non-RAG web baseline, using only web search results and a role prompt for grounding, while C removed both role and personalization cues to represent a pure web-only condition. Configuration D

introduced RAG with FAQs and policies, but without prompts or personalization. Hereafter, all configurations included the FAQs and policies within the RAG vector database. Building on this, E incorporated a role prompt to make retrieval role-aware, and F embedded user-specific adaptation directly within the retriever. Configuration G shifted context-aware guidance to the generation stage, applying a personalization prompt after retrieval, whereas H combined role context and embedded individualized generation for more contextualized outputs. Configuration I employed dual-layer prompting, adding both role and user-tailored prompts alongside retrieved data. Configuration J created redundant personalization, applying user context in both retrieval and generation, and K represented the fully integrated system, combining role input, retriever-level personalization, and redundant prompt-level personalization. Each configuration held decoding and 0-level temperature settings constant to isolate user-specific adaptation effects. These system configurations were designed to be compared in specific pairs. This pairing isolates the independent and interactive effects of role-framing, retrieval grounding, and personalization, as detailed in Table 3.

Table 3. Comparison pairs for evaluating experimental effects (excluding B)

Effect	Conf comparison	Description
Role effect (Web)	A-C	Impact of adding role framing to web query vs plain question.
Retrieval effect (RAG vs Web)	D-C and D-A	Effect of using VectorDB retrieval instead of web/no retrieval.
Role effect (RAG)	E-D	Effect of adding role framing in retrieval context.
Personalization in retrieval	F-D	Effect of embedding personalization inside VectorDB retrieval results.
Personalization in prompt	G-D	Effect of adding personalization in prompt text.
Personalization in prompt	F-G	Retrieval vs prompt source of personalization.
Interaction of role + Embedded personalization	H-F and H-E	Joint effect of role and embedded personalization.
Interaction of role + Prompt level personalization	I-G and I-E	Joint effect of role and prompt-based personalization.
Redundancy/Interference	J-F	Effect of redundant personalization (retrieval + prompt).
Full redundancy (Role + Retrieval + Prompt)	K-H and K-I	Effect of maximal redundancy of role and personalization sources.
Aggregate interaction (Role × Retrieval)	(E-D) vs (A-C)	Interaction of role and knowledge source.
Aggregate interaction (Role × Personalization)	(I-G) vs (H-F)	Interaction of role and personalization mode.
Aggregate interaction (Retrieval × Personalization)	(F-D) vs (G-D)	Interaction of knowledge source and personalization.
Aggregate interaction (Role × Retrieval × Personalization)	(K-I) vs (H-F)	Full triple interaction (redundancy effects).

3.3 *General system operation*

As shown in Figure 1, the system's operation begins with an authentication and session setup phase where the user enters university credentials. Upon receipt of a question, if the student's information is not in the cache, the personalization agent connects to the university portal, navigates to the logged-in student's degree information, scrapes it, and returns it. The system then converts the user's question and personalization data (as appropriate for the configuration) into the same vector space as the documents and performs a cosine similarity search between the query vector and the chunk vectors. The prompt assembly component then gathers the retrieved context, the original question, and (depending on the system configuration) a role prompt. Finally, this assembled input is given to the LLM for response generation.

4. Methodology

Three research questions guide the AiVisor experimentation. (RQ1) whether personalization improves semantic alignment and contextual relevance on question-answering tasks. (RQ2) asks whether personalization materially affects lexical fidelity. (RQ3) asks under what conditions personalization enhances or degrades performance within a RAG architecture. Based on these questions, two hypotheses were tested, as mentioned in the background section. H1 posits that personalization's effects are complex and metric-dependent, creating trade-offs between semantic, lexical, and reasoning quality, rather than uniform improvement. Specifically, H1 anticipates that in this lexically-skewed test, gains in reasoning and relevance may be coupled with penalties in semantic similarity metrics that rely on a single generic reference. H2 posits that personalization does not substantially alter lexical composition, indicating that any observed improvements occur at the level of meaning rather than expression.

AiVisor was evaluated using a pool of advising questions representing realistic student queries, each with a ground truth answer verified by a faculty advisor, a subset has been mentioned in Table 4. As established in the introduction, this question set was intentionally skewed toward lexical precision. The questions span fact recall, definitional clarity, constrained reasoning, and light synthesis. Each item was labeled as "context-sensitive" or "lexically strict" for subgroup analysis. Each system configuration generated responses for 100 simulated users, whose question selection was modeled using a Monte Carlo simulation in which each user randomly sampled between one and 12 questions from the question pool. Rather than fixing user profiles to specific demographic groups, this approach deliberately samples across a broad behavioral space, avoiding demographic bias and ensuring that system performance is evaluated across a wide range of user interaction patterns rather than a narrow-predefined population.

Table 4. Subset of advising questions

Advising questions
What courses are registered for the current academic term?
What steps do I need to take to address any unmet conditions for their degree requirements?
What are my rights regarding my educational records?
Describe the process I should follow if I believe there has been an error in my grade or if I wish to appeal a final course grade.

4.1 *Metrics*

AiVisor outputs are evaluated using eight complementary metrics, as shown in Table 5, that cover lexical, semantic and RAGAs-based grounding accuracy for every system-output, per item and using the same references. Lexical fidelity is measured by BLEU and ROUGE-L computed at the answer level against a single institutional reference. METEOR bridges lexical and semantic similarity by rewarding stemming and synonym matches. BERTScore measures embedding-based semantic alignment between generated and reference responses using contextual token embeddings. To assess grounding in retrieved evidence, we use four RAGAs measures: Faithfulness, Answer Relevancy, Context Precision, and Answer Correctness. Faithfulness quantifies whether the answer is supported by the retrieved passages. Answer Relevancy measures topical alignment to the user question. Context Precision captures the proportion of retrieved content that is necessary and used. Answer Correctness reflects factual and contextual accuracy. Since Systems A and C rely solely on web-based retrieval rather than a document-grounded RAG pipeline, RAGAs metrics are not reported for these configurations, as they require a controlled retrieval context derived from a fixed document corpus.

BLEU, ROUGE-L, and METEOR are computed with standard Python libraries (Natural Language Toolkit (NLTK) for BLEU and METEOR, rouge-score for ROUGE-L) with case preservation and punctuation retained. BERTScore is computed with the bert-base-uncased model in its default configuration and references are aligned token-wise with idf-weighting (inverse document frequency weighting) enabled. RAGAs metrics are produced using the opensource ragas

toolkit with deterministic settings. In particular, Faithfulness and Answer Relevancy are scored via an LLM evaluator interface with temperature 0.0 and a max tokens cap sufficient to cover full answers, while Context Precision and Answer Correctness consume the retrieved context windows produced by the system.

Table 5. Metric overview and interpretive focus

Metric	Type	Focus
BLEU	Lexical	n-gram overlap
ROUGE-L	Lexical	Longest common subsequence
METEOR	Lexical/Semantic	Stem, synonym, and paraphrase matching
BERTScore	Semantic	Meaning alignment via contextual embeddings
Faithfulness	Grounding	Consistency of answer with provided context
Answer Relevancy	Grounding	Relevance of answer to the user query
Context Precision	Grounding	Proportion of context that supports the answer
Answer Correctness	Grounding	Overall factual and contextual accuracy of the answer

To counter differences in metric scales and variances, all metric outputs are standardized across the system-by-item matrix before aggregation. Specifically, each metric column is z-normalized to zero mean and unit variance over all observations. Three primary composite indices were then constructed as the unweighted mean of these z-scores to capture performance families: Lexical_Index_z (BLEU, ROUGE-L); Semantic_Index_z (METEOR, BERTScore); and Reasoning_Index_z (Faithfulness, Answer Relevancy, Context Precision, Answer Correctness). Average ranks per metric and an overall mean rank are also computed as a scale-free check on ordering.

4.2 *Statistical analysis*

Given the complex, fractional-factorial nature of the system configurations (Table 2 and Table 3), where factors like personalization are nested with retrieval and role prompting, the primary analysis relies on Linear Mixed-effects Model (LMM) (statsmodels.MixedLM). This approach correctly handles the collinearity and repeated measures, allowing for the isolation of specific main effects and interactions of Role, Retrieval, and Personalization, while treating the question as a random intercept. As a secondary, exploratory analysis, user-adapted configurations are also compared to their nearest non-adapted baselines using paired t-tests (scipy.stats.ttest_rel) and Wilcoxon signed rank tests. Family-wise error across this set of metrics is controlled with the Holm-Bonferroni procedure at $\alpha = 0.05$. Effect sizes are expressed as Cohen's d for paired samples, accompanied by 95% confidence intervals estimated via percentile bootstrap with 10,000 resamples. Item-level win rates and pairwise win matrices are computed to summarize performance. A Pareto analysis is used to identify non-dominated solutions.

5. Results

The results reveal a complex trade-off with performance being highly dependent on the metric family (lexical, semantic, or reasoning). Simple paired t-tests, for instance, show contradictory results. Table 6 indicates a statistically significant decrease in BERTScore for user-adapted systems (Mean 0.841) versus non-adapted (Mean 0.848), even as METEOR significantly increased (Mean 0.361 vs 0.251). The primary LMM analysis clarifies this finding. The LMM, which can properly isolate interaction effects, revealed that while personalization factors significantly

improved Reasoning quality, they also had significant negative interactions with role-framing, which resulted in the semantic similarity loss observed in Table 6. This pattern is consistent with the "worst-case" evaluation design, where personalization's deviation from a single lexical reference is penalized by metrics like BERTScore. The RAGAs metrics confirm this trade-off, indicating that this semantic metric penalty did not degrade grounding; in fact, Answer Relevancy and Context Precision often improved. Correlation analysis in Figure 2 reinforces that semantic metrics correlate strongly with Answer Relevancy, while lexical overlap correlates weakly with Faithfulness.

Table 6. Statistical summary of personalized vs non-personalized performance. Values are mean ± SD. Δ reports the difference (personalized-non-personalized) with 95 percent bootstrap Confidence Intervals (CIs) where available

Metric	Mean (Pers)	Mean (Non-pers)	Δ (95% CI)	*p*-value
BLEU	0.106 ± 0.096	0.097 ± 0.127	0.009	0.0266
ROUGE-L	0.256 ± 0.128	0.242 ± 0.184	0.014	0.0124
METEOR	0.361 ± 0.195	0.251 ± 0.152	0.110	0.0000
BERTScore	0.841 ± 0.035	0.848 ± 0.031	-0.007 [-0.017, 0.020]	0.0000
Faithfulness	0.711 ± 0.411	0.655 ± 0.344	0.056 [-0.011, 0.233]	0.0135

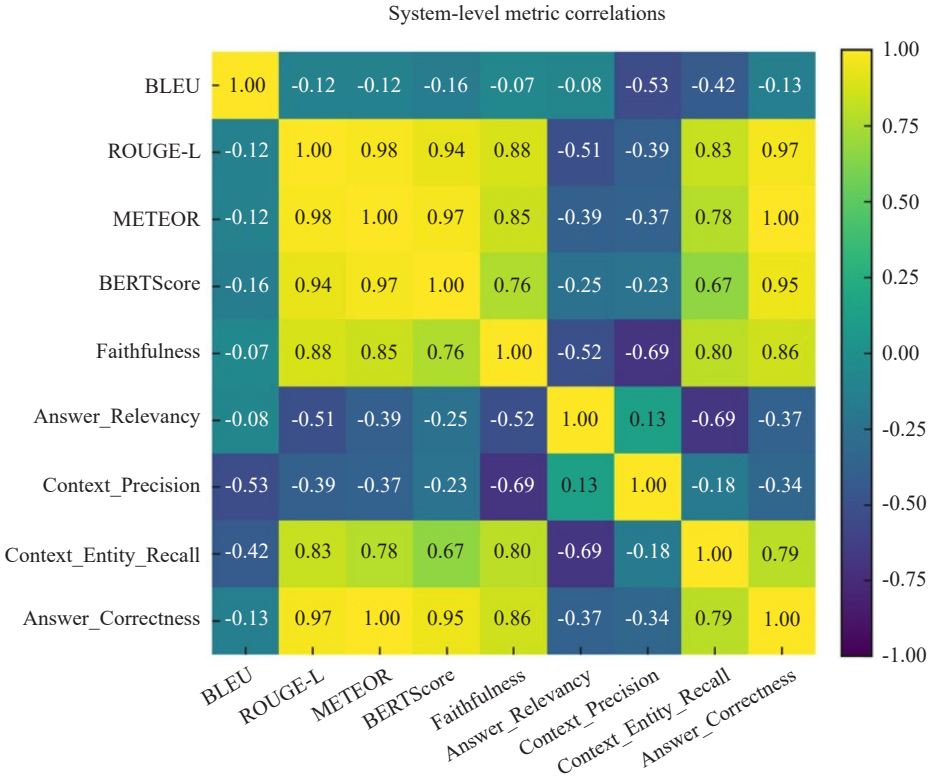


Figure 2. System-level means illustrating correlation strength across all eight metrics

The fractional-factorial design required a LMM as the primary analysis. This approach resolves the contradictions seen in simpler paired comparisons, such as the conflicting BERTScore and METEOR results in Table 6. The LMM results, presented in Table 7, reveal that personalization does not produce uniform gains but rather a critical, metric-

dependent trade-off. The most significant findings are not the main effects, but the complex interactions. Specifically, combining personalization with role-framing significantly improved reasoning quality. This is evident in the Reasoning_*z* results for the Role: PersPrompt interaction (coef: 0.041, $p = 0.028$) and the three-way interaction (coef: 0.044, $p = 0.018$). However, these exact same interactions had a significant negative impact on Semantic_*z* (Role: PersPrompt coef: -0.114, $p = 0.000$; three-way coef: -0.082, $p = 0.002$). This semantic penalty, also reflected in the strong negative main effect for PersRetrieval (coef: -0.116, $p = 0.000$), is the expected outcome of the conservative, lexically-skewed experimental design. The semantic metrics are demonstrably penalizing personalization for correctly deviating from the single, generic ground truth answer, thereby confirming the observed loss is a methodological artifact even as the reasoning-focused metrics confirm a simultaneous and beneficial improvement in answer quality.

Table 7. LMM results for fixed effects. Non-estimable (collinear) factors like C (retrieval, sum) were automatically dropped by the model

Metric	Effect	Coef	Std_err	*z*_value	*p*_value
LexicalIndex_*z*	Main effect of role framing	0.026	0.024	1.093	0.274
	Main effect of personalization in retrieval	-0.025	0.024	-1.037	0.300
	Main effect of personalization in prompt text	-0.050	0.024	-2.120	0.034
	Role moderates retrieval-embedded personalization	0.062	0.024	2.623	0.009
	Role moderates prompt personalization	0.042	0.024	1.753	0.080
	Dual personalization redundancy/interference	0.058	0.024	2.439	0.015
	Role moderates redundancy	-0.025	0.024	-1.039	0.299
Semantic_*z*	Main effect of role framing	0.057	0.027	2.140	0.032
	Main effect of personalization in retrieval	-0.116	0.027	-4.336	0.000
	Main effect of personalization in prompt text	-0.046	0.027	-1.723	0.085
	Role moderates retrieval-embedded personalization	0.006	0.027	0.240	0.810
	Role moderates prompt personalization	-0.114	0.027	-4.258	0.000
	Dual personalization redundancy/interference	-0.114	0.027	-0.531	0.595
	Role moderates redundancy	-0.082	0.027	-3.089	0.002
Reasoning_*z*	Main effect of role framing	-0.022	0.019	-1.160	0.246
	Main effect of personalization in retrieval	-0.069	0.019	-3.672	0.000
	Main effect of personalization in prompt text	-0.106	0.019	-5.687	0.000
	Role moderates retrieval-embedded personalization	-0.054	0.019	-2.888	0.004
	Role moderates prompt personalization	0.041	0.019	2.204	0.028
	Dual personalization redundancy/interference	0.005	0.019	0.266	0.790
	Role moderates redundancy	0.044	0.019	2.356	0.018

The visualization of standardized metrics provides the foundational evidence for the trade-off. Figure 3 summarizes the performance of all configurations by *z*-score across all eight metrics. This heatmap visually confirms the source of

the high composite scores seen in the personalized systems (K, J, I). These systems are characterized by a broad band of positive *z*-scores (yellow/green) across the grounding and reasoning metrics, demonstrating consistent outperformance in factual alignment and contextual relevance. In contrast, non-personalized baselines (A, C, D) cluster heavily in negative *z*-score territory (purple), highlighting the effectiveness of the added personalization factors in shifting performance.

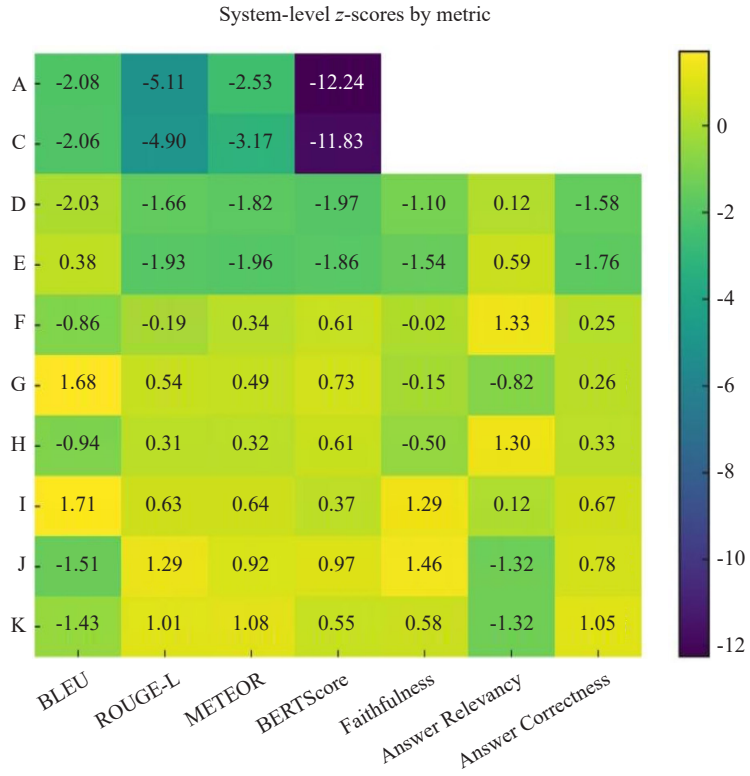
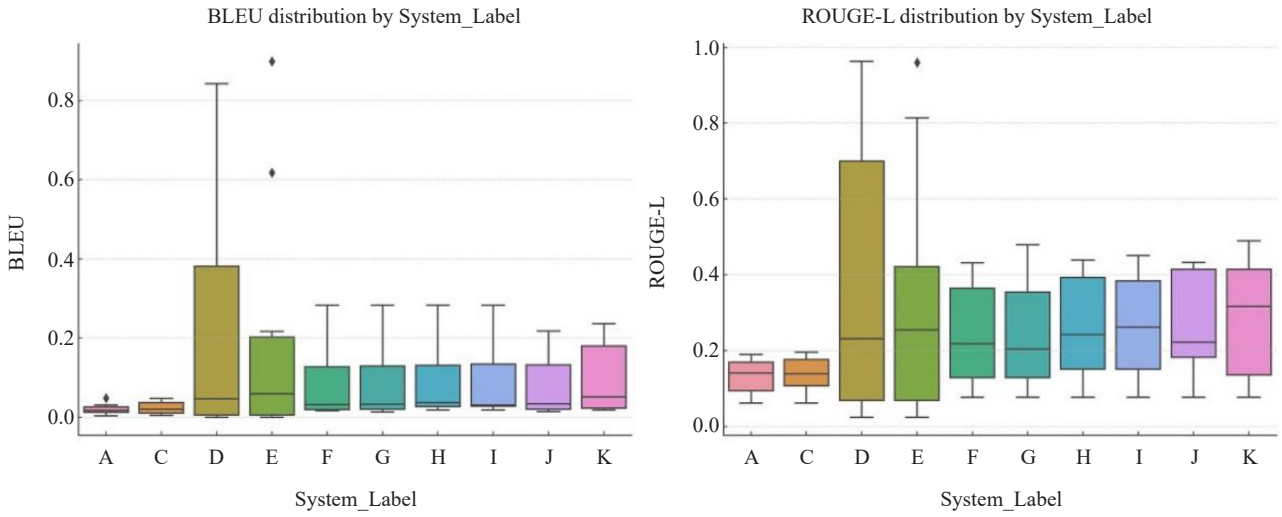


Figure 3. Standardized (*z*-score) system performance by metric. Positive values indicate above-average performance for that metric relative to other systems; blanks indicate unavailable metrics



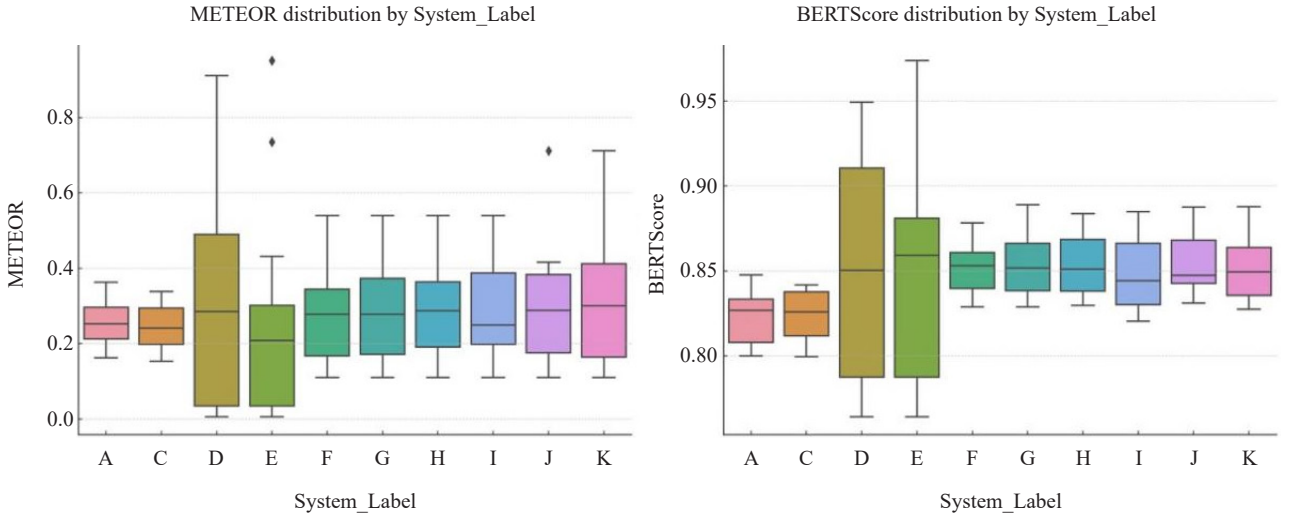


Figure 4. Distribution of BLEU, ROUGE-L, METEOR, and BERTScore across systems. Box plots illustrate within-metric variance and outliers, providing a baseline view of lexical and semantic dispersion prior to normalization and system-level aggregation

Visualization of raw data provides the necessary context for interpreting normalized results. Figure 4 illustrates the distribution of the four classic similarity metrics in all system configurations. This figure highlights the dramatic differences in the metric scale and dispersion (e.g. the tight concentration of the BERTScore versus the wide variance in METEOR and ROUGE-L), reinforcing the methodological need to apply the normalization of the *z*-score before aggregation or direct comparison.

5.1 *Composite score and ranking*

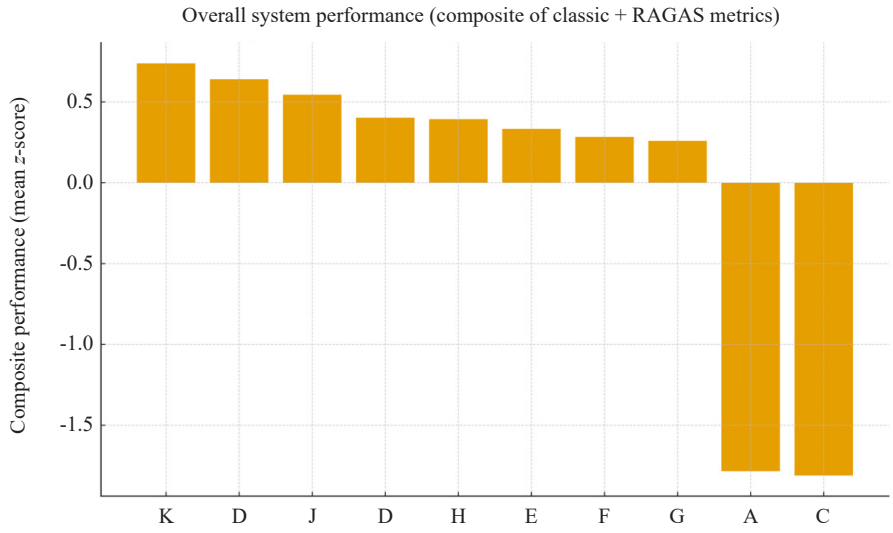


Figure 5. Overall system (composite) performance

The composite *z*-score, which aggregates all eight metrics, provides a high-level view of performance. As shown in Figure 5, fully integrated personalized systems (K, J, and I) achieve the highest composite scores. However, this aggregated view obscures the critical trade-off identified in the LMM analysis (Table 7). This high composite score is not the result of uniform improvement. Rather, the strong gains in reasoning and grounding metrics (Reasoning_*z*) were

substantial enough to outweigh the significant semantic loss (Semantic_z). This trade-off is a direct consequence of the lexically-skewed evaluation design. System K's top ranking is thus attributable to achieving the most balanced trade-off, rather than a uniform improvement across all individual metrics. This finding is reinforced by the rank analysis, where personalized variants consistently occupy lower mean ranks due to their strength in these non-lexical, reasoning-based categories.

Rank analysis corroborates the composite score findings. Each metric column contributes to a rank matrix whose row-wise average defines the mean rank per system, where lower ranks correspond to better overall performance. Personalized variants consistently occupy lower mean ranks; however, this is not due to uniform improvement. This strong average ranking is driven by their high performance in reasoning and grounding metrics, which successfully compensates for their poor ranking in semantic similarity (as shown in Table 7). This trade-off is clearly illustrated in the Pareto analysis in Figure 6. This figure situates the systems on an efficiency frontier, showing that System D and System K each represent an optimal trade-off. System D maximizes lexical quality (METEOR), while System K maximizes grounding quality (Faithfulness).

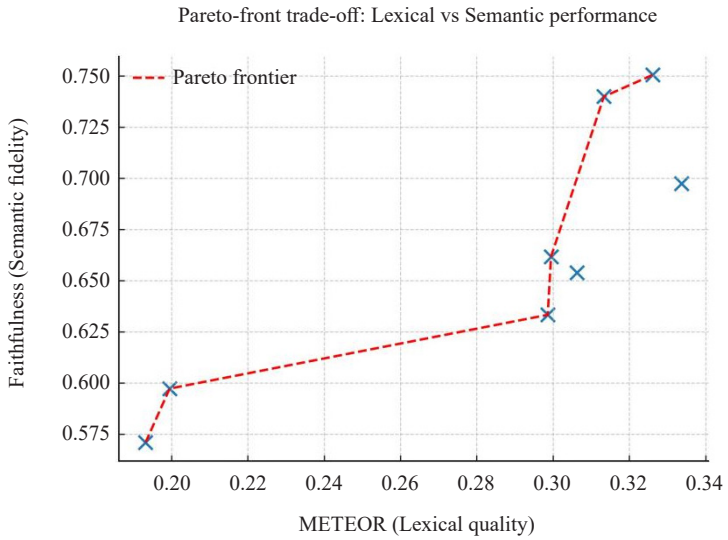


Figure 6. Pareto-front trade-off between lexical quality (METEOR) and grounding fidelity (Faithfulness) at the system level. The dashed line indicates the Pareto frontier; systems on the frontier each represent an optimal trade-off between these two competing objectives

5.2 *System comparison and pareto structure*

The detailed system comparison clarifies how personalization interacts with role prompting and retrieval conditioning. System D, the pure VectorDB baseline, achieves strong lexical alignment because its retrieved context preserves canonical institutional language, but fails to adapt to user intent. System E, which introduces a role prompt without personalization, provides moderate improvements in topical relevance and Faithfulness. System F, which applies user-specific adaptation only in retrieval, substantially increases Context Precision by emphasizing passages that align with user profiles, yet may underperform on abstract reasoning questions. System I, which uses prompt-level user-tailored responses with a role prompt, exemplifies the central trade-off identified in the LMM analysis (Table 7): it achieves strong reasoning gains but incurs a significant semantic penalty in BERTScore. System J, which applies individualized generation at retrieval and prompt levels but omits explicit role prompting, strengthens Faithfulness but introduces occasional scope drift when roles overlap semantically. System K, combining all three mechanisms (role prompting, retriever-level adaptation, and prompt-level adaptation), achieves the highest composite score (Figure 5) and the lowest mean rank.

To empirically isolate the contribution of personalization redundancy in System K from the effect of individual placement decisions, a diagnostic discordance metric X is introduced:

$$X = Z(\text{Answer Relevancy}) - Z(\text{BERTScore})$$

where Z(Answer Relevancy) denotes the z-normalized RAGAs Answer Relevancy score, a reference-free measure assessing the degree to which the generated response addresses the posed question and Z(BERTScore) denotes the z-normalized BERTScore F1, a reference-dependent metric quantifying token-level semantic similarity between the generated answer and the ground truth using contextual embeddings. Z-normalization is applied prior to subtraction to neutralize scale discrepancies arising from the fundamentally different scoring distributions of the two metrics, yielding a dimensionless measure of directional discordance between question relevance and reference faithfulness. A negative X value indicates that the system exhibits stronger reference-faithful behavior than question-relevant generation, while a positive value indicates the reverse.

The diagnostic metric was computed for Systems H and I, selected as the two nearest two-layer counterparts to K by virtue of sharing two of K's three personalization mechanisms, alongside K itself, enabling a controlled partial decomposition of redundancy contributions (Table 8).

The marginal contribution of adding prompt-level personalization to a system already incorporating role input and retrieval personalization is computed as $X(\text{K}) - X(\text{H}) = -2.56$, while the marginal contribution of adding retrieval-level personalization to a system already incorporating role input and role prompt is computed as $X(\text{K}) - X(\text{I}) = -1.62$. The larger magnitude of the former suggests that retrieval-side personalization exerts a stronger influence on reference-relevance within the fully integrated configuration, while prompt-level personalization provides a meaningful but comparatively smaller contribution.

Table 8. Diagnostic discordance metric $X = Z(\text{Answer Relevancy}) - Z(\text{BERTScore})$ for two-layer and three-layer personalization configurations

System	Z(Answer Relevancy)	Z(BERTScore)	X
H	1.30	0.61	0.69
I	0.12	0.37	-0.25
K	-1.32	0.55	0.110

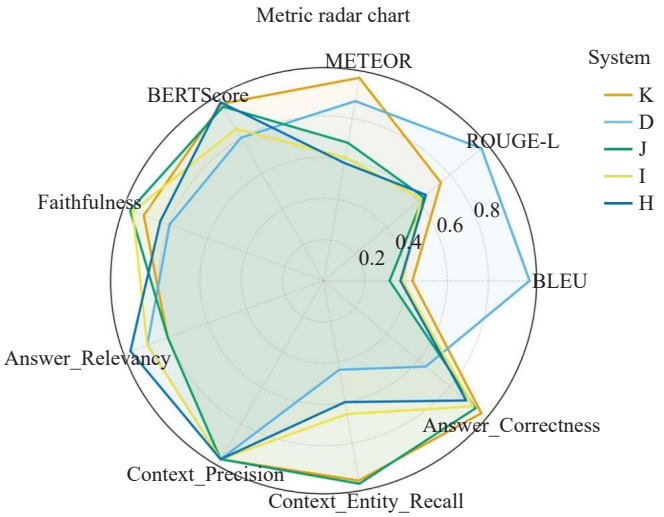


Figure 7. Multi-metric radar profile for the top-5 systems by composite normalized score

The multi-dimensional nature of the trade-off is further shown in the radar chart in Figure 7. This visualization

clearly shows that System K, the composite leader, achieves its top ranking not by dominating all metrics, but by balancing performance across different families. Specifically, System K demonstrates peak scores in grounding metrics such as Context Precision and Faithfulness, but its semantic score (BERTScore) is often lower than the raw RAG baseline (System D). This visual evidence reinforces the LMM finding that the personalization strategy yields an optimal blend of reasoning fidelity and contextual relevance, even when penalized by semantic metrics.

5.3 *Significance and robustness*

Personalization is most beneficial when questions are based on who the user is or what rules apply to their situation, as this information limits the reasoning space and improves accuracy. For generic fact lookup questions, personalization adds qualifiers that change the surface form without altering the core meaning. This deviation from the single lexical ground truth is what causes the semantic penalty (e.g. in Semantic_*z*) observed in Table 7, even while reasoning quality (e.g. Reasoning_*z*) simultaneously improves. An informal review of the generated responses suggested that there was no increase in hallucination. However, few instances of excessively narrow responses were observed when redundant cues narrowed the scope too aggressively; selectively disabling personalization on lexically strict or compliance-sensitive items could mitigate this risk.

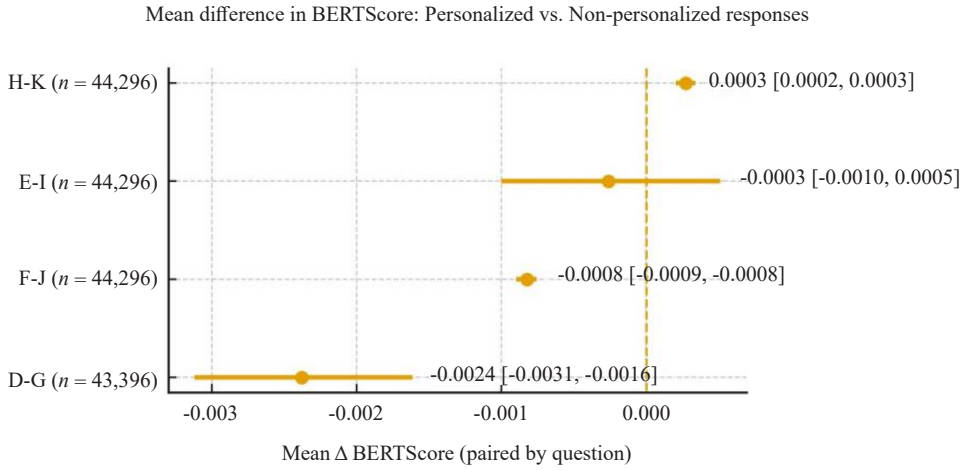
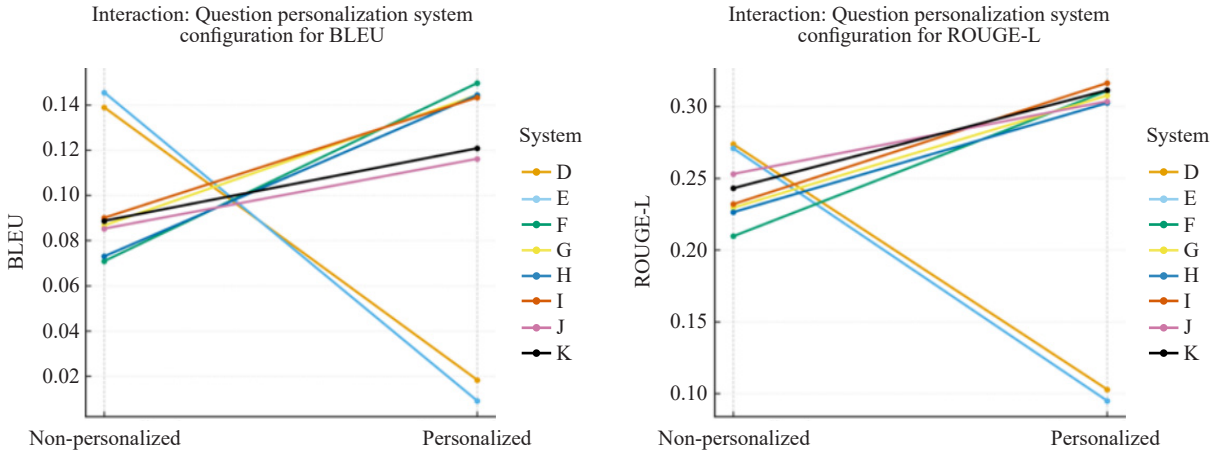
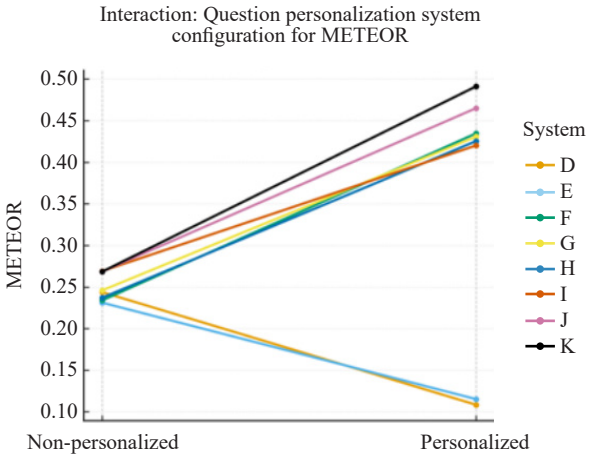
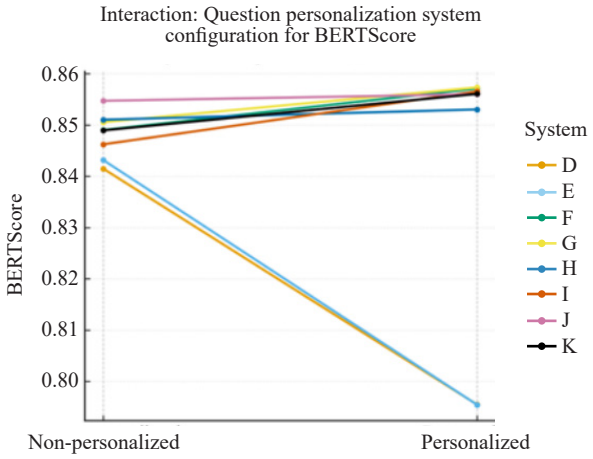
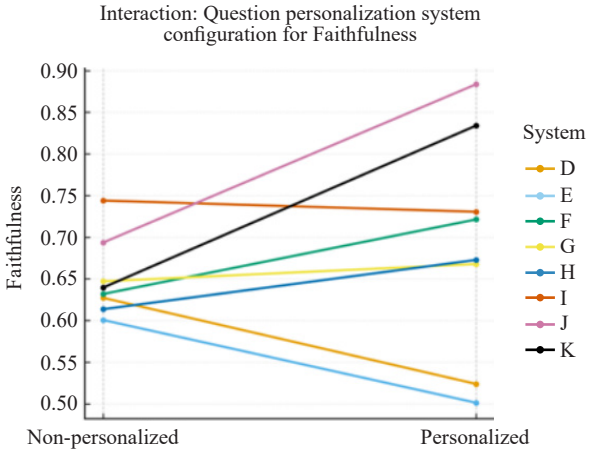


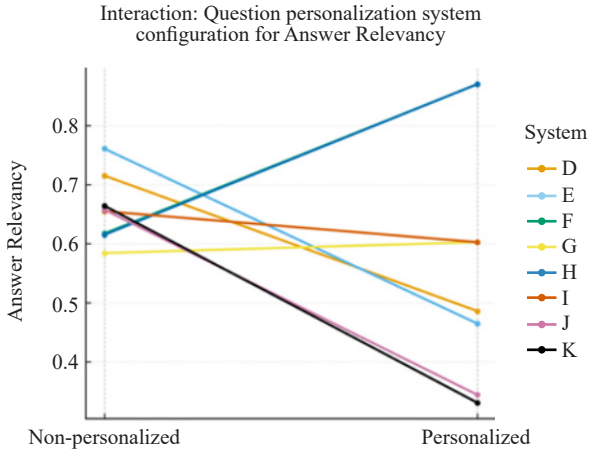
Figure 8. Forest plot of personalization effect sizes (Δ BERTScore) for paired systems, with 95 percent bootstrap CIs. Each bar shows the mean delta (personalized-baseline) with 95% bootstrap confidence intervals for paired systems D-G, E-I, F-J, and H-K

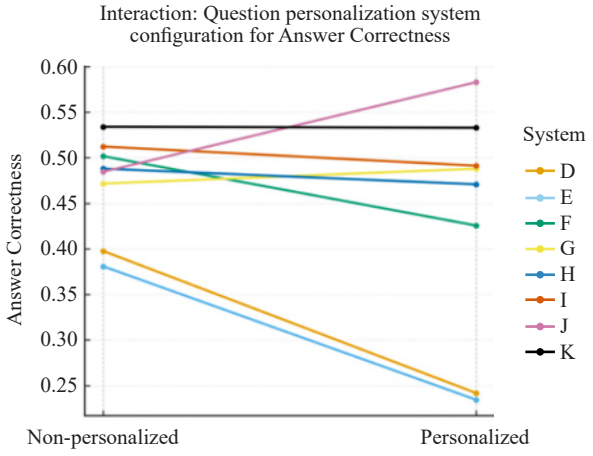


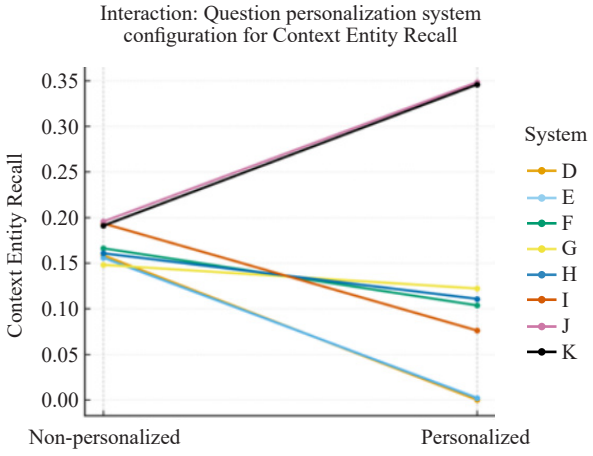

Interaction: Question personalization system configuration for METEOR
METEOR
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
Non-personalized
Personalized
System
D
E
F
G
H
I
J
K


Interaction: Question personalization system configuration for BERTScore
BERTScore
0.86
0.85
0.84
0.83
0.82
0.81
0.80
Non-personalized
Personalized
System
D
E
F
G
H
I
J
K


Interaction: Question personalization system configuration for Faithfulness
Faithfulness
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
Non-personalized
Personalized
System
D
E
F
G
H
I
J
K


Interaction: Question personalization system configuration for Answer Relevancy
Answer Relevancy
0.8
0.7
0.6
0.5
0.4
Non-personalized
Personalized
System
D
E
F
G
H
I
J
K


Interaction: Question personalization system configuration for Answer Correctness
Answer Correctness
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
Non-personalized
Personalized
System
D
E
F
G
H
I
J
K


Interaction: Question personalization system configuration for Context Entity Recall
Context Entity Recall
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Non-personalized
Personalized
System
D
E
F
G
H
I
J
K

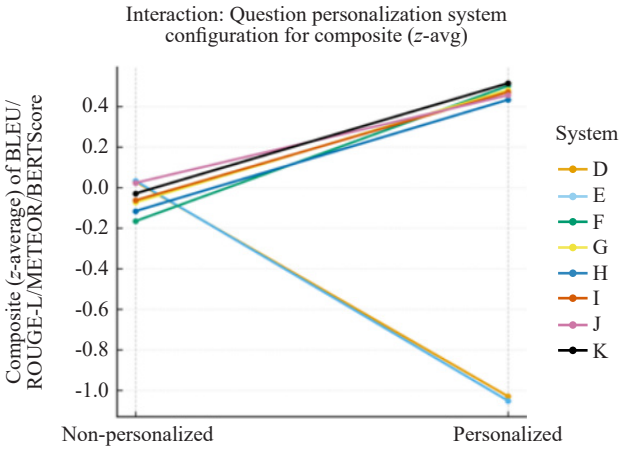


Figure 9. Interaction plots by metric. Each panel shows system-level traces (A-K) across non-personalized and personalized questions with the legend placed at right within each panel. Panels include BLEU, ROUGE-L, METEOR, BERTScore, Faithfulness, Answer Relevancy, Answer Correctness, Context Entity Recall, and Composite (*z*-average of BLEU/ROUGE-L/METEOR/BERTScore)

The Linear Mixed-Effects Model (Table 7) provides the primary statistical analysis for this complex design, and its findings are supported by other results. For example, the "semantic loss" identified by the LMM is also visually apparent in the simpler paired comparisons. Figure 8 visualizes the mean difference in BERTScore, clearly showing a statistically significant negative shift for most personalized configurations (such as D-G and E-I). This plot confirms the semantic penalty that the LMM analysis more precisely isolates and explains. Furthermore, Figure 9 supports Hypothesis H1 by illustrating a clear personalization-by-question effect, where the performance of different system configurations (the plotted lines) diverge based on the question type.

5.4 *Prompting, retrieval, and personalization effects*

Three experimental factors shape performance: role prompting, retrieval grounding, and the stage at which personalization is applied. The LMM results (Table 7) clarify the distinct impact of each. Role prompting, for instance, had a significant main effect on semantic quality (coef: 0.057). Personalization embedded in the retriever (PersRetrieval) had a strong negative main effect on both semantic (coef: -0.116) and reasoning (coef: -0.069) quality. Personalization applied at the prompt level (PersPrompt) also had a strong negative main effect on reasoning (coef: -0.106). However, the most important finding is how these factors interact. The combination of Role and PersPrompt created a significant trade-off: it improved reasoning (coef: 0.041) while simultaneously harming semantic quality (coef: -0.114). The radar profile in Figure 7 illustrates these trade-offs, showing how System K (all factors) excels in grounding metrics like Faithfulness but not in semantic metrics like BERTScore, relative to other systems.

This complex trade-off is visualized in the interaction plots shown in Figure 9. These plots illustrate the performance of each system configuration across different metrics. The plots for Faithfulness and Answer Correctness clearly show an upward trend for the personalized, integrated systems (e.g., K, I, J). Conversely, the plot for BERTScore shows a distinct downward or flat trend for these same systems, visually confirming the semantic loss penalty identified in the LMM analysis. This figure demonstrates that the strong gains in reasoning and grounding metrics were substantial enough to offset the semantic penalty, leading to an overall upward trend in the Composite_*z* score.

6. Discussion and implications

The AiVisor experiment provides critical evidence of the complex trade-offs in agentic personalization, moving beyond the field's assumption of uniform benefit. The primary LMM analysis revealed that personalization's effects are not uniformly positive; instead, it creates a trade-off between reasoning and semantic similarity. The analysis found significant negative interactions between role-framing and personalization factors on semantic metrics (e.g.,

BERTScore: Mean (Pers) 0.841 vs. Mean (Non-Pers) 0.848), alongside positive interactions on reasoning metrics. This result, which may seem counter-intuitive, is directly explained by the experimental design. The test was systematically disadvantaged against personalization, using a question set "skewed toward lexical precision" and evaluating against a single, generic ground truth. Personalization succeeds by deviating from this generic reference to be more user-specific. Therefore, the observed semantic loss is a methodological artifact of standard metrics penalizing user-specific adaptation for correctly doing its job. The fact that reasoning and grounding scores improved simultaneously (e.g., RAGAs Faithfulness: Mean (Pers) 0.711 vs. Mean (Non-Pers) 0.655), confirms that this deviation was beneficial, not harmful. These results contribute methodologically by demonstrating that multi-locus userspecific adaptation produces complex interactions. Theoretically, the findings support two conclusions. First, standard evaluation benchmarks are insufficient for user-tailored Q & A, as they fail to account for beneficial deviation from a generic reference. Second, personalization acts as a critical constraint on both evidence selection at retrieval and expression at generation.

Compared with prior advising and RAG-based systems, AiVisor offers several notable distinctions. First, unlike prior advising studies such as Advisely and FIT-Advisor, which primarily demonstrate feasibility or localized usefulness, AiVisor isolates personalization as the experimental variable and provides evidence of exactly when and where it helps. This offers a more mechanistic understanding of personalization effects than simply demonstrating that an AI advisor can answer questions, as it explains the conditions under which performance improves. Second, compared with agentic RAG work such as [34], which showed that an agentic architecture can outperform retrieval-only baselines on RAGAs metrics, AiVisor extends this line of inquiry by demonstrating that performance depends on where personalization is injected: retrieval-only personalization improves context targeting, prompt-level personalization improves reasoning under role guidance, and the fully redundant configuration (System K) yields the best overall trade-off by aligning retrieval, role framing, and generation simultaneously. Third, AiVisor employs a more comprehensive evaluation framework than much of the adjacent literature, assessing systems across lexical, semantic, and grounding metrics together and resolving apparent contradictions using a Linear Mixed-Effects Model. This is important because a simpler analysis would risk incorrectly concluding that personalization harms answer quality due to the BERTScore decrease, whereas the full analysis reveals a more nuanced picture on the dimensions most relevant to advising: contextual relevance, faithfulness to retrieved evidence, and answer correctness.

The performance advantage observed in the fully integrated configurations appears to stem from their ability to align three decision layers: role framing, retrieval conditioning, and response generation. This alignment allows the system to select more relevant evidence, reason under student-specific constraints, and produce answers that are more contextually grounded than non-personalized or single-layer RAG approaches. System K is the clearest illustration of this, achieving the highest composite performance and lowest mean rank by converting personalization from a superficial prompt feature into a coordinated retrieval-and-generation strategy. The diagnostic discordance analysis (Table 8) further supports this interpretation. The differential marginal contributions of retrieval-side and prompt-level personalization confirm that K's advantage stems from coordinated redundancy across all three layers rather than from any single placement acting in isolation, a finding consistent with the positive three-way interaction identified in the LMM analysis (Reasoning_z coef: 0.044, $p = 0.018$).

6.1 *Personalization redundancy: systems I, J, and K*

Comparison among systems I, J, and K isolates the impact of redundancy and role prompting. System I (prompt-level user-tailored responses with a role) exemplifies the central trade-off identified in the LMM analysis: it produces significant reasoning gains while incurring a significant semantic penalty. System J (retrieval-level user-specific adaptation) improves evidence relevance but may drift in pragmatic scope without role guidance. System K integrates both and adds role prompting, achieving a synergistic effect specifically on reasoning, as shown by the positive three-way interaction for Reasoning_z (coef: 0.044, $p = 0.018$). The redundancy observed in K appears to form a feedback alignment between retrieval bias and generation framing, leading to superior Context Precision and Faithfulness gains (Faithfulness: 0.711 vs. 0.655, $p = 0.0135$). This dual-conditioning mechanism, or redundant personalization, can be understood as a lightweight form of constraint harmonization similar in concept to bi-encoder feedback loops where acts to align the output expression with the specific retrieved context. This is computationally necessary in high-stakes domains like advising, where failure to apply specific constraints (e.g., student major) leads to high-cost algorithmic errors [5].

Redundant personalization introduces risks of over-constraining the model, where user-specific cues may propagate excessively through both retriever and generator. To mitigate this, a selective mechanism should modulate redundancy by detecting when retrieval already provides highly focused context. In those cases, prompt-level personalization can be reduced to avoid excessively narrow or privacy-sensitive phrasing. Conversely, when retrieval context is diffuse, prompt personalization should remain active to maintain relevance and user alignment.

6.2 *Practical implications and implementation guidance*

The practical implications are direct for system design. Role prompting should be enabled by default to maintain tone and domain scope consistency. VectorDB grounding is essential for factual scaffolding, as originally proposed in retrieval-augmented generation frameworks [21]. Personalization in retrieval should be applied when user attributes directly affect relevance, which can improve Context Precision and Answer Relevancy. Prompt-level personalization is a key factor in the reasoning-semantic trade-off, as the LMM analysis revealed. It should be used to improve reasoning quality, where the analysis showed a positive interaction. However, designers must be aware that it will likely incur a semantic penalty (a BERTScore loss) when evaluated against generic references (Role: PersPrompt Reasoning_z coef: 0.041, $p = 0.028$; Semantic_z coef: -0.114, $p < 0.001$). When both mechanisms operate together, this redundancy should be selectively managed to avoid propagating sensitive user attributes. In operational deployments, the hierarchy observed in this study offers practical guidance. System I is recommended when retrieval cannot be modified. System J is preferable when retrieval must be user-aware but roles are homogeneous. System K is the right choice when both retriever and generator can share secure state information. System K thus represents the configuration with the best composite performance, achieving the most effective balance between reasoning gains and semantic penalties.

Beyond advising, the methodological pattern observed here generalizes to other retrieval-augmented tasks where personalization or context adaptation plays a role, including educational tutoring, clinical decision support, and customer interaction systems. By demonstrating the complex trade-offs between reasoning gains and semantic penalties under conservative conditions, AiVisor establishes a baseline and a methodological framework for future research in robust and transparent personalization for agentic AI.

7. Limitations and future work

The generalizability of these trade-offs is constrained by the intentionally conservative test environment built on twelve lexically-strict questions. We strategically minimized the evaluation scope to isolate and maximize the detection of the semantic penalty, but full validation requires replication on a larger, multi-turn, human-judged suite to confirm external validity. Although paired tests, effect sizes, sensitivity checks, and re-weighting help mitigate this limitation, a larger question suite will be required for full validation. The task mix is intentionally skewed toward generic, lexically strict prompts to create a conservative, worst-case setting; results should therefore be re-examined on multi-turn and domain-specific evaluations. The chosen four-metric panel captures core aspects of lexical and semantic quality but remains incomplete, as human judgments, calibration scores, and instructionfollowing measures warrant inclusion. System generality is also limited by reliance on a single base model family and a fixed prompt template set, motivating replication across alternative architectures and personalization schemas.

The current chunking configuration, which reduces retrieval precision and weakens personalization signals, is a key limitation. This factor, combined with the intentionally lexically-skewed question set, created the conservative test environment. The LMM analysis confirmed the outcome of this design: improvements were observed in RAGAs-based reasoning metrics, but these occurred alongside a significant penalty in semantic metrics (like BERTScore). This penalty is interpreted as a direct methodological artifact of evaluating user-specific deviations against a single, generic reference. Future work will extend the question set, refine retrieval granularity, and incorporate human assessments of appropriateness and usefulness. This is essential for addressing the ethical challenges of algorithmic bias and transparency [14] by confirming that the algorithmic reasoning gain translates into trustworthy and fair guidance from a human perspective.

8. Conclusion

AiVisor demonstrates that agentic personalization produces complex, metric-dependent trade-offs, not simple uniform gains, even under conditions unfavorable to personalization. The Linear Mixed-Effects Model analysis, necessitated by the fractional-factorial design, isolated a key conflict: personalization factors (especially when interacting with a role prompt) significantly improved reasoning quality while simultaneously incurring a statistically significant penalty on semantic similarity metrics like BERTScore (BERTScore: 0.841 vs. 0.848, $p < 0.0001$; METEOR: 0.361 vs. 0.251, $p < 0.0001$). This semantic loss can therefore be interpreted as a methodological artifact of the conservative, personalization evaluation, where lexically-oriented prompts and coarse chunking penalize deviations from a generic reference. Among tested configurations, System K-combining all factorsachieved the highest composite score, driven by strong gains in reasoning and grounding that outweighed the semantic metric penalty, with Faithfulness improving from 0.655 to 0.711 ($p = 0.0135$). This supports the hypothesis that the semantic loss was not a true loss of quality. These outcomes constitute a conservative validation of personalization, showing it yields measurable benefits in the correctness of an answer, even as it challenges the method used to measure semantic similarity.

Beyond empirical findings, AiVisor advances personalization research through methodological rigor and transparency. The evaluation framework integrates lexical and semantic similarity metrics with advanced statistical testing, moving beyond small-scale pilots and perception studies to provide system-level evidence of personalization effects. The analysis employed per-metric *z*-scoring and composite aggregation, while a Linear Mixed-Effects Model was used as the primary statistical test. This mixed-model approach was necessary to de-conflict the complex, nested factors of the experimental design, thereby providing the most robust framework for analyzing personalization's true effect. Future work will confirm the dominance of reasoning factors by calculating a Weighted Composite Score using the LMM coefficients as weights.

Evaluation under lexical stringency strengthens external validity by testing personalization in settings where it is least advantaged, conditions under which BLEU and ROUGE-L penalties are most likely; the multi-metric, normalized design addresses comparability across metrics and reduces the risk of selective reporting, while item-level accountability highlights which questions benefit from the reasoning gains and which incur lexical or semantic penalties. This granularity enables the development of selective mechanisms that determine when personalization should be applied or withheld, supporting responsible deployment.

Overall, AiVisor serves both as a prototype platform for academic advising and as an experimental testbed for studying personalization as an independent variable. The findings establish that personalization's effects are not uniformly positive, but rather create a complex trade-off between reasoning gains and semantic penalties (when measured against a lexical reference), providing an empirical foundation for cautious, data-driven personalization in production assistants. Future research should expand task coverage, incorporate human ratings of appropriateness and usefulness, extend to multiturn workflows, and formalize selective mechanisms that balance the benefits of individualized generation with reliability in real-world applications.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1] Thüs D, Malone S, Brünken R. Exploring generative AI in higher education: a RAG system to enhance student engagement with scientific literature. *Frontiers in Psychology*. 2024; 15: 1474892. Available from: https://doi.org/10.3389/fpsyg.2024.1474892.

[2] Lim JJY, Zhang-Li D, Yu J, Cong X, He Y, Liu Z, et al. Learning in context: personalizing educational content with large language models to enhance student learning. *arXiv:2509.15068*. 2025. Available from: https://doi.org/10.48550/arXiv.2509.15068.

[3] Hasan T, Bunescu R. A survey of affective recommender systems: modeling attitudes, emotions, and moods for

personalization. *arXiv:2508.20289*. 2025. Available from: https://doi.org/10.48550/arXiv.2508.20289.
[4] Li J, Elgarf M, Waleed A, Salam H. Beyond one-size-fits-all: a survey of personalized affective computing in human-agent interaction. *arXiv:2304.00377*. 2023. Available from: https://doi.org/10.48550/arXiv.2304.00377.
[5] Li X, Jia P, Xu D, Wen Y, Zhang Y, Zhang W, et al. A survey of personalization: from RAG to agent. *ACM Transactions on Information Systems*. 2026; 4(4): 93. Available from: https://doi.org/10.1145/3802586.
[6] Liu J, Qiu Z, Li Z, Dai Q, Yu W, Zhu J, et al. A survey of personalized large language models: progress and future directions. *arXiv:2502.11528*. 2025. Available from: https://doi.org/10.48550/arXiv.2502.11528.
[7] Chandra S, Verma S, Lim WM, Kumar S, Donthu N. Personalization in personalized marketing: trends and ways forward. *Psychology & Marketing*. 2022; 39(8): 1529-1562. Available from: https://doi.org/10.1002/mar.21670.
[8] Dudy S, Bedrick S, Webber B. Refocusing on relevance: personalization in NLG. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p.5190-5202. Available from: https://doi.org/10.18653/v1/2021.emnlp-main.421.
[9] Bunt H, Petukhova V. Semantic and pragmatic precision in conversational AI systems. *Frontiers in Artificial Intelligence*. 2023; 6: 896729. Available from: https://doi.org/10.3389/frai.2023.896729.
[10] Blömker J, Albrecht CM. Reevaluating personalization in AI-powered service chatbots: a study on identity matching via few-shot learning. *Computers in Human Behavior: Artificial Humans*. 2025; 3: 100126. Available from: https://doi.org/10.1016/j.chbah.2025.100126.
[11] Kim J, Yang Y. Few-shot personalization of LLMs with mis-aligned responses. *arXiv:2406.18678*. 2024. Available from: https://doi.org/10.48550/arXiv.2406.18678.
[12] Akiba D, Fraboni MC. AI-supported academic advising: exploring ChatGPT's current state and future potential toward student empowerment. *Education Sciences*. 2023; 13(9): 885. Available from: https://doi.org/10.3390/educsci13090885.
[13] Antico C, Giordano S, Koyuturk C, Ognibene D. Unimib assistant: designing a student-friendly RAG-based chatbot for all their needs. *arXiv:2411.19554*. 2024. Available from: https://doi.org/10.48550/arXiv.2411.19554.
[14] Siddique AB, Maqbool MH, Taywade K, Foroosh H. Personalizing task-oriented dialog systems via zero-shot generalizable reward function. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*. New York, NY, USA: ACM; 2022. p.1787-1797. Available from: https://doi.org/10.1145/3511808.3557417.
[15] Casaca JA, Miguel LP. The influence of personalization on consumer satisfaction: trends and challenges. In: *Data-Driven Marketing for Strategic Success*. Hershey, PA: IGI Global Scientific Publishing; 2024. p.256-292. Available from: https://doi.org/10.4018/979-8-3693-3455-3.ch010.
[16] Tarifi N, Bakhsh R. The effectiveness of personalized marketing strategies on consumer engagement and loyalty (Evidence from KSA). *International Journal of Research Studies in Publishing*. 2024; 5(56): 283-302. Available from: https://doi.org/10.52133/ijrsp.v5.56.11.
[17] Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ questions for machine comprehension of text. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, TX: ACL; 2016. p.2383-2392. Available from: https://doi.org/10.18653/v1/D16-1264.
[18] Bajaj P, Campos D, Craswell N, Deng L, Gao J, Liu X, et al. MS MARCO: a human generated machine reading comprehension dataset. *arXiv:1611.09268*. 2018. Available from: https://doi.org/10.48550/arXiv.1611.09268.
[19] Kwiatkowski T, Palomaki J, Redfield O, Collins M, Jones L, Potts C, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*. 2019; 7: 452-466. Available from: https://doi.org/10.1162/tacl_a_00276.
[20] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*. 2023; 55(12): 248. Available from: https://doi.org/10.1145/3571730.
[21] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NY, United States: Curran Associates Inc.; 2020. p.9459-9474.
[22] Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey. *arXiv:2312.10997*. 2023. Available from: https://doi.org/10.48550/arXiv.2312.10997.
[23] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA: ACL; 2002. p.311-318. Available from: https://doi.org/10.3115/1073083.1073135.
[24] Lin CY. ROUGE: a package for automatic evaluation of summaries. In: *Text Summarization Branches Out*.

Barcelona, Spain: ACL; 2004. p.74-81.
[25] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: ACL; 2005. p.65-72.
[26] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. Addis Ababa; 2020.
[27] Thottoli MM, Alruqaishi BH, Soosaimanickam A. Robo academic advisor: can chatbots and artificial intelligence replace human interaction? *Contemporary Educational Technology*. 2024; 16(1): ep485. Available from: https://doi.org/10.30935/cedtech/13948.
[28] Lin Y, Yu Z. A bibliometric analysis of artificial intelligence chatbots in educational contexts. *Interactive Technology and Smart Education*. 2024; 21(2): 189-213. Available from: https://doi.org/10.1108/ITSE-12-2022-0165.
[29] Bilquise G, Ibrahim S, Shaalan K. Bilingual AI-driven chatbot for academic advising. *International Journal of Advanced Computer Science and Applications*. 2022; 13(8): 50-57. Available from: https://doi.org/10.14569/IJACSA.2022.0130808.
[30] Dawood M. Assessing the effectiveness of chatbots in providing personalized academic advising and support to higher education students: a narrative literature review. *Studies in Technology Enhanced Learning*. 2024; 4(1): 1-12. Available from: https://doi.org/10.21428/8c225f6e.7140f8f4.
[31] Abdelhamid S, Bangura J, Shah S. Advisely: AI-powered academic advising using large language models (LLMs). In: *New Perspectives in Science Education*. Florence, Italy; 2025.
[32] Soomro AA, Khan MH, Umar M, Khan S, Ali O. AI-driven academic advising in higher education: leveraging intelligent systems to personalize student support, improve retention, and optimize career pathways. *The Critical Review of Social Sciences Studies*. 2025; 3(2): 229-248. Available from: https://doi.org/10.59075/vy3v7k17.
[33] Luong H, Luong K. A chatbot-based academic advising model for students in information technology: a case study. *Saudi Journal of Engineering and Technology*. 2025; 10(3): 93-100. Available from: https://doi.org/10.36348/sjet.2025.v10i03.007.
[34] Tamascelli M, Bunch O, Fowler B, Taeb M, Cohen A. Academic advising chatbot powered with AI agent. In: *Proceedings of the 2025 ACM Southeast Conference (ACMSE 2025)*. Cape Girardeau, MO, USA: ACM; 2025. p.195-202. Available from: https://doi.org/10.1145/3696673.3723065.
[35] Reid M, Savinov N, Teplyashin D, Lepikhin D, Lillicrap T, Alayrac JB, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*. 2024. Available from: https://doi.org/10.48550/arXiv.2403.05530.
[36] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*. 2025; 43(2): 1-55. Available from: https://doi.org/10.1145/3703155.