Constraint-Aware Retrieval-Augmented Generation for Business-Event Venue Recommendation: A Component Ablation Study

(Author names and affiliation are supplied in OJS submission metadata, per the CommIT double-blind policy.)

Abstract

Selecting venues for Meetings, Incentives, Conferences, and Exhibitions (MICE) events in Indonesia remains manual and slow, while keyword search cannot interpret multi-criteria natural-language requests. This study develops Vento AI, an Advanced Retrieval-Augmented Generation (RAG) chatbot recommending hotel venues over a knowledge base of 1,741 Indonesian venues across 36 provinces and 142 cities and regencies, and asks which retrieval techniques actually improve recommendation quality. Multi-Query Retrieval (MQR), Metadata Filtering (MF), and Hypothetical Document Embeddings (HyDE), fused through Reciprocal Rank Fusion, are varied in a full 2x2x2 factorial ablation of eight configurations answering the same 40 ground-truth queries, giving 320 runs measured with deterministic metrics and tested using the Wilcoxon signed-rank test under Holm correction. Contrary to the assumption that stacking techniques is better, the complete pipeline is not the best. Metadata filtering alone reaches a Context Precision@8 of 1.000 at a Constraint Violation Rate of 0%, against 0.247 and 50.0% for the semantic baseline and 0.366 and 20.0% for the complete pipeline. Rerunning the ablation on a second generator reproduces the inference exactly: all four MF-bearing configurations significantly outperform the baseline in both models, and none of the three without MF does. A second set of 40 numeral-free requests bounds the finding, the MF advantage shrinking to 0.300 yet remaining significant. Evaluation with 38 valid respondents returns Technology Acceptance Model means of 4.41 to 4.47 and a System Usability Scale mean of 77.7, above the normative 68. For constrained entity retrieval, filtering placement matters more than technique count.

Keywords: retrieval-augmented generation; metadata filtering; venue recommendation; ablation study; cross-model replication

I. INTRODUCTION

The Meetings, Incentives, Conferences, and Exhibitions (MICE) industry depends on rapid, accurate matching of an event's requirements to suitable hotel venues. In practice this matching is still done manually and across fragmented channels: planners compile requirements such as city, delegate capacity, star class, package scheme, and per-participant budget, and then contact venues one by one. Conventional keyword search offers little relief because it matches surface terms rather than interpreting a multi-criteria natural-language request such as "a full-day meeting in Bandung for 300 people under IDR 500,000 per participant." The result is a slow, expertise-heavy process that is difficult for non-specialists to navigate. The MICE sector is undergoing a rapid, technology-driven transformation whose adoption readiness remains unevenly mapped [1], [2], [3], [4], while venue selection specifically rests on multi-criteria assessment frameworks [5].

Retrieval-Augmented Generation (RAG) grounds a large language model (LLM) in an external, updatable knowledge base, reducing hallucination by conditioning generation on retrieved evidence [6]. Beyond the naive retrieve-then-read pipeline, a family of Advanced RAG techniques refines the retrieval and generation stages [7], [8]. Multi-Query Retrieval (MQR) expands a query into several paraphrases to widen coverage [9]. Hypothetical Document Embeddings (HyDE) generates a hypothetical ideal answer and searches with its embedding [10] and Metadata Filtering (MF) applies hard structured constraints to the candidate set, an approach shown to sharpen retrieval when structured evidence is combined with dense vectors [11]. When several retrieval paths are used, their ranked lists are commonly merged with Reciprocal Rank Fusion (RRF), which rewards items ranked highly across lists [12]. A widely held working assumption is that combining more of these techniques yields better retrieval.

That assumption is rarely tested by isolating each technique's contribution on a constrained entity-retrieval task, and even less so for Indonesian-language requests in the MICE domain. Much prior work evaluates RAG with LLM-graded metrics [13] on open-ended question answering, where fluency and faithfulness dominate, and current evaluation surveys still centre on those targets [14]. Component-level studies of RAG pipelines do exist [15], [16], but they tune open-domain question answering rather than constrained entity retrieval. Meanwhile, LLMs are increasingly applied to recommendation and ranking , [17], [18] , and RAG has been explored for tourism and hospitality recommendation [19], [20], [21], while hotel or venue recommenders have traditionally relied on ratings,

reviews, and collaborative filtering [22]. Venue recommendation, however, is a constrained task whose correctness can be verified deterministically against structured data, since whether a recommended venue actually satisfies the city, capacity, class, and budget constraints is a matter of fact rather than of judgment. This gap motivates a controlled ablation that measures each technique's true effect with objective, reproducible metrics rather than LLM judgments.

This study makes four contributions. First, it develops Vento AI, an Advanced RAG venue-recommendation chatbot deployed over a domain knowledge base of 1,741 Indonesian hotel venues spanning 36 provinces and 142 cities and regencies. Second, it evaluates the constituent retrieval techniques through a full factorial (2x2x2 = 8) ablation on 40 ground-truth queries, using deterministic metrics and the Wilcoxon signed-rank test with Holm correction [23]. Third, it reports a counter-intuitive finding and establishes its robustness on two fronts: the whole ablation is rerun on a second, independently developed generator, and a second query set written entirely in numeral-free natural language bounds the conditions under which the finding holds. Fourth, it triangulates the machine evaluation with a user-acceptance study using the Technology Acceptance Model (TAM) and the System Usability Scale (SUS).

The study is guided by two research questions. RQ1: which composition of the retrieval-refinement techniques (MQR, MF, and HyDE fused through RRF) produces the best venue-recommendation quality, measured against ground truth, and does that answer hold when the generator and the linguistic form of the request are varied? RQ2: how do MICE practitioners accept and rate the usability of the resulting system?

II. METHODS

This is a development-research study with a quantitative approach that combines building a software artifact with controlled empirical evaluation. It follows two complementary evaluation paths: an ablation experiment that measures each retrieval technique's contribution to machine quality over ground-truth queries, and a user evaluation that assesses acceptance and usability of the deployed prototype. The two paths answer different questions, namely which technique composition is best by machine measures and whether the system is accepted and usable by practitioners, and they are not substitutes for each other.

A. Knowledge Base and Data

The knowledge base comprises 1,741 hotel venues in Indonesia across 36 provinces and 142 cities and regencies, covering 91 hotel groups, compiled by the author from more than 14 years of MICE-industry practice together with public sources. Each venue carries identity attributes (name, hotel group, star class), location, meeting-room capacity, seven package-price attributes, facility attributes, and a free-text description. All prices are quoted per participant (per pax), following Indonesian corporate-quotation practice, and represent two service schemes. The non-residential meeting scheme has three durations: half-day (about four hours with one coffee break), full-day (about eight hours with one meal and two coffee breaks), and full-board (about twelve hours with two meals and two coffee breaks), each covering meeting-room use and consumption without an overnight stay. The full-board residential scheme bundles overnight accommodation with a full-board meeting into a single per-participant price in single, twin, and triple room-occupancy variants, with an extra-bed attribute representing quadruple occupancy. Table 1 maps the seven price attributes to these schemes.

Table 1. MICE meeting-package schemes and their per-participant price attributes.

Scheme	Package	Description (per participant)	Price attribute
Non-residential meeting	Half-day	~4 hours, 1 coffee break; room + consumption, no overnight stay	halfday
	Full-day	~8 hours, 1 meal and 2 coffee breaks	fullday *
	Full-board	~12 hours, 2 meals and 2 coffee breaks	fullboard
Full-board residential	Single	Full-board meeting + overnight stay, 1 guest per room	residential-single
	Twin / Triple	As above, 2 / 3 guests per room	residential-twin / triple
	Extra bed	Additional bed (quadruple occupancy)	extrabed

*Source: authors (2026). * Full-day is complete for all 1,741 venues and is used as the budget-filter field.*

The dataset was frozen as a research snapshot, identified by a SHA-256 checksum, so that all experiments read identical data. Preprocessing served three ends. First, categorical attributes used as filtering metadata were normalized: venue type, originally 21 mixed labels with 311 blank entries (17.9%), was consolidated into six standard categories, and the concept tag was reduced from 54 mixed values to six. Second, price fields were curated: units were normalized, duplicate entities were removed, and a documented median imputation by brand and star class was applied where a package price was missing. Third, location was rebuilt rather than accepted as supplied. Province had originally been taken verbatim from the source field, which admitted city-province pairs that are administratively impossible; each venue's province was therefore re-derived from its street address and validated against a 313-entry reference table of Indonesian regencies and cities. This corrected the city or province of 85 rows and raised readable coverage from 32 to 36 provinces and from 131 to 142 cities and regencies without adding or removing a single venue. All 1,741 rows now carry a street address, a map link, an administrative city or regency, and a recorded provenance for the location assignment. Colloquial search names such as Bali and Jakarta were deliberately retained so that the query sets remain matchable. All 1,741 rows were preserved and the original values were kept as an audit trail, with normalized columns added separately. After curation the knowledge base is complete on every key filtering attribute: capacity and full-day price are populated for all 1,741 venues, half-day and full-board prices likewise reach full coverage, and facility data covers 99.4% of rows. Public rating, available for only 3.8% of rows, was excluded from the pipeline. One residual issue is reported rather than suppressed: 365 venues share a street address with a venue recorded in a different city, and these rows are flagged openly in the curation notes as pending field verification.

Before the knowledge base was put into service, a diagnostic learnability check verified that its structured attributes carry consistent signal rather than noise. Two baseline classifiers were trained to predict a venue's star class from five non-price features (capacity, facility count, venue type, concept tag, and province) under a stratified 80:20 split with five-fold cross-validation. On the curated knowledge base, Random Forest reached 0.848 test accuracy (macro-F1 0.628; cross-validation 0.830 ± 0.024) and Gradient Boosting 0.848 (macro-F1 0.630; cross-validation 0.841 ± 0.013), well above the 51.6% majority-class baseline. These figures are lower than on the pre-curation snapshot, which is expected: correcting the province field widened regional diversity and made the classification task harder. The conclusion is unchanged, in that the structured attributes remain informative enough to serve both as filtering metadata and as venue-profile representation. This check sits outside the recommendation pipeline.

B. System Architecture

The architecture separates two stages. Offline, and only once, the 1,741 venue records are turned into one text document per venue, embedded with a multilingual Sentence-BERT model, and written to a vector index that stores the structured attributes alongside each vector; no component of the recommendation pipeline re-reads the

raw records at query time. Online, and for every user query, the pipeline operates as follows. An Indonesian-language user query is processed through three parallel retrieval paths (MQR, MF, and HyDE) that all read that one shared vector index; the ranked lists from active paths are merged with RRF (constant k = 60), assembled into a bounded, reordered context that mitigates the Lost-in-the-Middle effect [25], and passed to the generator, which composes a recommendation with justifications. A system instruction requires the generator to answer only from the attached context and to present prices as indicative references, suppressing hallucination in the black-box LLM setting [26]. Fig. 1 illustrates both stages, with a concrete example on the online side (query Q05): a request for a full-day meeting venue in Bandung for about 300 participants within a IDR 450,000 per-participant budget, showing the paraphrases produced by MQR, the hard filter extracted by MF, and the hypothetical ideal document drafted by HyDE, together with the fused result and the final recommendation.

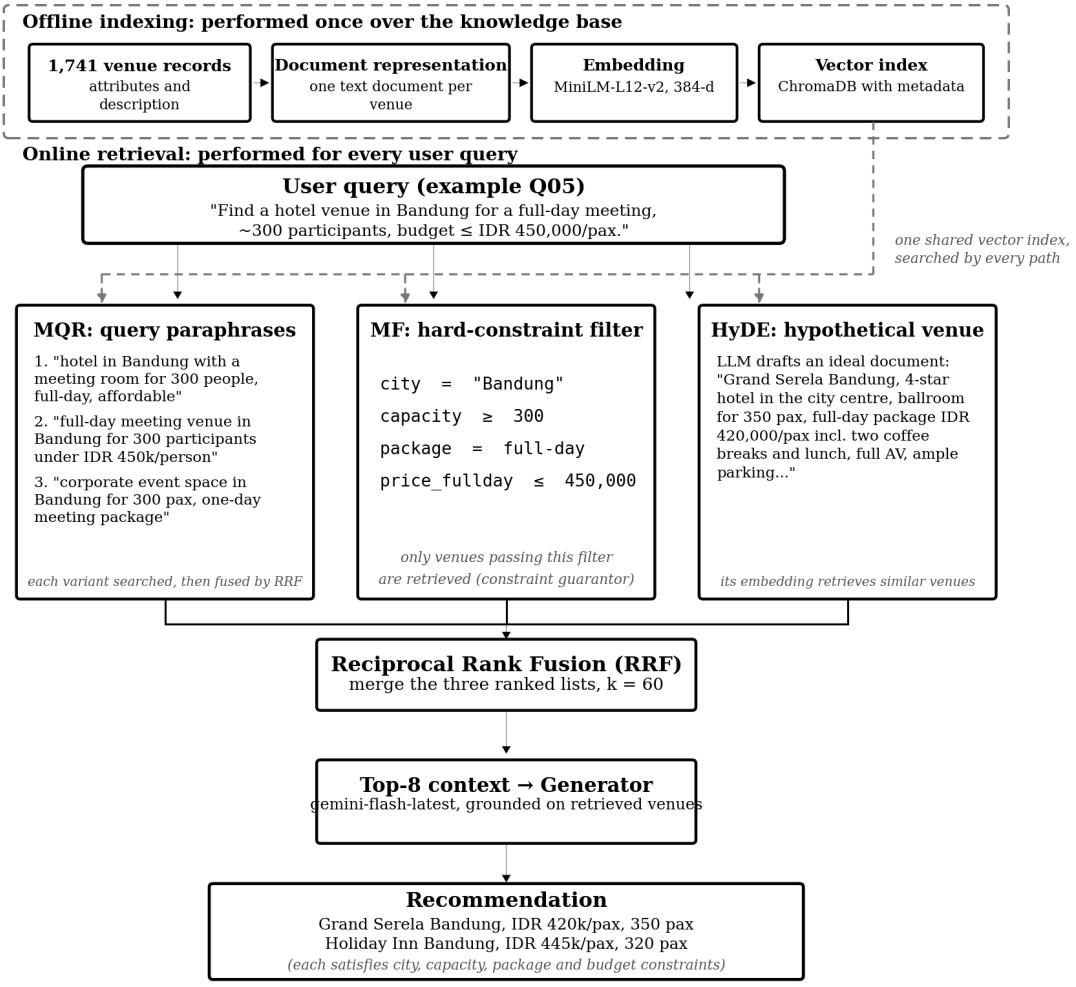


Fig. 1. The Vento AI architecture with a worked example for query Q05. The upper band is the offline indexing stage, performed once over the knowledge base; the lower part is the online retrieval performed for every query, showing the concrete output of MQR, MF, and HyDE before fusion and generation.

Technology choices balanced research-design fit, Indonesian-language support, and compute cost. The generator is a black-box LLM accessed via API, a setting in which retrieval is prepended to a model whose weights are never updated [26]; the primary experiments use the endpoint gemini-3.6-flash, accessed August 2026, with generation temperature 0.2 for answers, 0.7 for query paraphrases, and 0.5 for HyDE documents. Embeddings

use paraphrase-multilingual-MiniLM-L12-v2 from SentenceTransformers [27], a 384-dimensional multilingual model chosen for Indonesian coverage and inference cost. Vectors are stored in ChromaDB with cosine distance, and metadata filtering is executed by the store's native where clause over city, star class, capacity, and the budget ceiling on the requested package. Retrieval draws 20 candidates per active path; the ranked lists are merged with RRF (constant k = 60), assembled into a bounded, reordered context of eight documents that mitigates the Lost-in-the-Middle effect [25], and the prompt instructs the model to answer only from the attached context and to present prices as indicative references.

Vento AI is deployed as a conversational assistant on a commercial event-planning platform in Indonesia; the platform is not named here in order to preserve double-blind review and will be identified in the camera-ready version with the operator's written permission. Fig. 2 shows the deployed system answering a live natural-language query (a half-day meeting in Bogor): the assistant returns ranked venue recommendations that include indicative per-participant price, capacity, and star class, together with request-booking actions, directly within the planner's workflow.

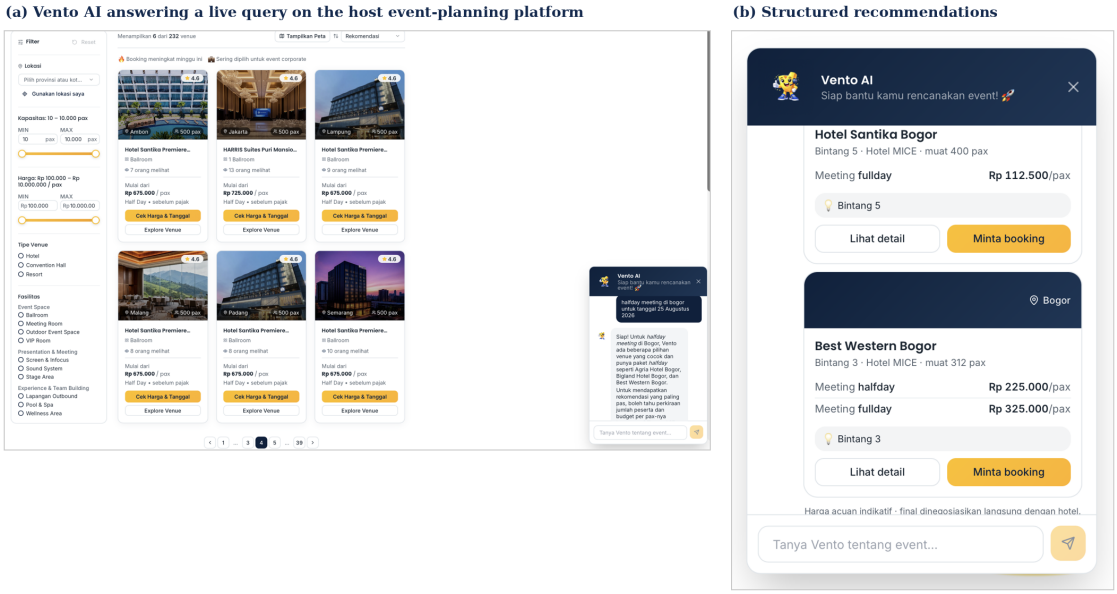


Fig. 2. The deployed Vento AI: (a) answering a live query on the host event-planning platform; (b) the structured venue recommendations returned.

Source: authors' deployed prototype (2026).

C. Ablation Design

The three techniques were treated as binary factors (on/off) in a full 2x2x2 = 8 factorial design, giving eight configurations from the naive-RAG equivalent (semantic retrieval only) to the complete pipeline, as illustrated in Fig. 3. Each configuration answered the same set of 40 Indonesian queries (listed in Table 2) drawn from real MICE request patterns and spanning large and small cities, capacities from tens to thousands of participants, all three package schemes, and queries with and without budget or accommodation constraints, including hard queries satisfiable by only a few venues. Every query carries a ground-truth list of venues that satisfy all its criteria, verified directly against the structured data. Running all configurations over all queries produced 8 x 40 = 320 pipeline executions.

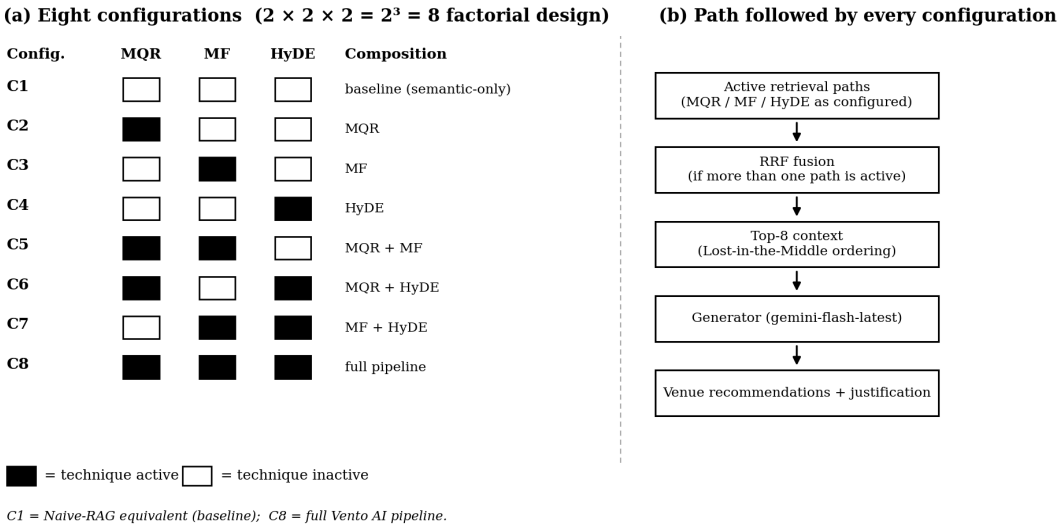


Fig. 3. The eight ablation configurations (2x2x2 factorial) and the shared pipeline each configuration follows.

Table 2. The first query set: 40 explicitly constrained ground-truth queries.

#	Query specification	#	Query specification
Q01	Jakarta, full-day, ~500 pax, ≤500k	Q21	Malang, full-board, ~200 pax, ≤1.0M, stay
Q02	Jakarta, half-day, ~200 pax, ≤350k	Q22	Bekasi, full-day, ~200 pax, ≤450k
Q03	Jakarta, ≥5★, full-board, ~1000 pax, ≤2.5M, stay	Q23	Tangerang, full-day, ~300 pax, ≤500k
Q04	Jakarta, ≥4★, full-day, ~300 pax, ≤600k, stay	Q24	Palembang, full-day, ~200 pax, ≤450k
Q05	Bandung, full-day, ~300 pax, ≤450k	Q25	Balikpapan, full-day, ~200 pax, ≤500k
Q06	Bandung, ≥4★, full-board, ~150 pax, ≤1.2M, stay	Q26	Padang, full-day, ~200 pax, ≤450k
Q07	Bandung, half-day, ~1000 pax, ≤400k	Q27	Solo, full-board, ~300 pax, ≤1.0M, stay
Q08	Bali, ≥5★, full-day, ~500 pax, ≤800k	Q28	Pekanbaru, half-day, ~200 pax, ≤300k
Q09	Bali, full-board, ~500 pax, ≤1.5M, stay	Q29	Banda Aceh, full-day, ~150 pax, ≤600k
Q10	Surabaya, full-day, ~400 pax, ≤500k	Q30	Manado, full-day, ~150 pax, ≤500k
Q11	Surabaya, ≥4★, half-day, ~150 pax, ≤400k	Q31	Any city, ≥5★, full-day, ~2000 pax, ≤700k
Q12	Yogyakarta, full-board, ~300 pax, ≤1.1M, stay	Q32	Any city, ≥5★, full-board, ~1500 pax, ≤1.5M, stay
Q13	Yogyakarta, full-day, ~500 pax, ≤450k	Q33	Any city, full-day, ~3000 pax, ≤800k
Q14	Bogor, full-board, ~200 pax, ≤1.0M, stay	Q34	Any city, ≥4★, half-day, ~100 pax, ≤200k
Q15	Bogor, full-day, ~500 pax, ≤500k	Q35	Jakarta, ≥5★, full-day, ~2000 pax, ≤1.8M
Q16	Semarang, full-day, ~300 pax, ≤450k	Q36	Bali, ≥5★, full-board, ~1000 pax, ≤3.3M, stay
Q17	Makassar, full-day, ~300 pax, ≤500k	Q37	Yogyakarta, ≥5★, full-day, ~800 pax, ≤1.6M
Q18	Medan, half-day, ~200 pax, ≤350k	Q38	Bandung, ≥5★, full-board, ~500 pax, ≤1.8M, stay
Q19	Batam, full-board, ~300 pax, ≤1.2M, stay	Q39	Semarang, ≥5★, full-day, ~500 pax, ≤600k
Q20	Lombok, full-day, ~200 pax, ≤600k	Q40	Makassar, ≥5★, full-day, ~800 pax, ≤650k

The queries were issued in Indonesian; the specifications below are English summaries. Notation: pax = participants; ≤ = maximum budget per participant (k = thousand, M = million IDR); stay = overnight accommodation required; ≥N★ = minimum star class. This set is used for the ablation in Table 3 and for the cross-model replication in Table 4; the second, numeral-free set described in Section II-E is reported separately in Table 5.

Source: authors (2026).

Quality was measured with five objective metrics computed deterministically against each query's ground truth, without any LLM judge: Context Precision@8 (fraction of correct documents among the eight retrieved context items), Recommendation Precision@5 (fraction of correct final recommendations), Hit Rate (fraction of queries with at least one correct recommendation), Mean Reciprocal Rank (MRR), and Constraint Violation Rate (CVR, the percentage of queries whose recommendations violate a stated constraint). Ground-truth metrics were chosen over LLM grading because this is a constrained entity-retrieval task whose ground truth is directly verifiable, making measurement objective, exactly reproducible, and free of the variability, bias, and cost that recent work documents in LLM-as-a-judge grading [13], [29], [36]. Each configuration's score is the mean over 40 queries. For significance testing, each query was summarized into one composite score defined as the unweighted mean of Context Precision@8, Recommendation Precision@5, Hit Rate, MRR, and constraint compliance expressed as one minus CVR. Paired differences were tested with the Wilcoxon signed-rank test [28] at the 5% level; because every configuration is compared against a common reference within each test family, p values were adjusted with the Holm step-down correction [38]. A non-parametric test was chosen because per-query scores are not assumed normal and per-cell sample sizes are small. The pipeline and all experiments were implemented in Python: the prototype uses SentenceTransformers for embeddings, ChromaDB for vector storage and metadata filtering, and a hosted LLM API for generation, while the ablation runs and statistical tests were executed in Google Colab using pandas and SciPy.

D. Cross-Model Replication

A single generator leaves open the possibility that the configuration ranking reflects the behavior of one language model rather than a property of the retrieval architecture. To separate the two, the complete eight-configuration ablation was rerun in full on a second, independently developed generator, gpt-5.6-luna, over the same frozen knowledge-base snapshot and the same 40 ground-truth queries. Every other element was held identical: embedding model, candidates per retrieval path, context size, RRF constant, metadata-filter fields, prompt template, and the definition of ground truth. Because the metrics and the query set are unchanged, the two runs are directly comparable value by value, and the comparison is reported alongside the primary results rather than as a separate complementary check. Two configurations serve as an internal validity control within this design. The semantic baseline C1 and the metadata-filtering configuration C3 invoke no language model anywhere on the retrieval path, so their retrieval-side metrics must be invariant to the generator; the remaining six configurations call the model to build paraphrases or hypothetical documents and are therefore expected to move.

E. Open-Language Query Set

The first query set states its constraints explicitly, and its ground truth is derived from the same constraint axes that metadata filtering operates on. Part of the MF advantage is therefore structural by construction. To establish the boundary of the finding rather than to defend it, a second set of 40 requests was written in natural Indonesian containing no numerals at all, so that city, star class, capacity, and budget are at most implied by word choice. Ground truth for this set was not derived from the filter: a prose relevance rubric was written and locked for each query before the system was run, and the qualifying venue set was then materialized from that rubric. The constraints supplied to metadata filtering were extracted from the query text by a rule-based parser that mirrors the deployed system, rather than being provided directly. The requests were grouped by the kind of constraint left implicit, with one English rendering of each group given here: implied location, as in "our team wants an offsite in a city with cool air, not too far from Jakarta, reachable by road" (10 queries); facility and atmosphere, as in "somewhere quiet and away from the crowds, suitable for training that needs concentration" (10); implied budget, as in "our budget is tight, find the most economical one-day meeting venue" (8); implied size, as in "a single-division event, with quite a lot of participants" (7); and narrative requests, as in "our client is a mining company and wants the event held in Kalimantan, at a good hotel with an overnight stay" (5). This set was run on C1, C2, and C3, the three configurations that isolate the baseline, query expansion, and filtering, in order to keep the number of generation calls within quota.

F. User Evaluation

The target population is Indonesian event-ecosystem practitioners; a purposive sample was drawn with a target of at least 30 valid respondents. Inclusion and exclusion criteria were fixed before analysis. A response was retained only if the respondent works within the event ecosystem, that is, as an event-organizer owner, staff member, or freelancer, as a vendor or partner, or as a representative of a company that commissions event organizers. The response also had to come from someone who had actually used the prototype, contain no blank questionnaire item, and pass a consistency check that rejects uniformly identical ratings given to positively and negatively worded items alike. Respondents used the deployed Vento AI for a real task scenario and then completed an online questionnaire of 20 closed five-point Likert items plus one open question, either independently at their own workplace or immediately after a guided walkthrough session. The first ten items measure TAM constructs adapted from Davis [24], namely perceived usefulness (PU, 3 items), perceived ease of use (PEOU, 3 items), attitude (2 items), and behavioral intention (BI, 2 items), and the next ten are an Indonesian translation of the SUS that preserves the original item order and polarity , an instrument whose validity for conversational interfaces has been confirmed on contemporary systems [25]. Item validity was assessed with corrected item-total correlation and reliability with Cronbach's alpha (threshold 0.70). SUS scores were computed with the standard rule and compared against the normative average of 68 [26].

$$SUS = 2.5 \times [\ \Sigma \text{ odd } (s - 1) + \Sigma \text{ even } (5 - s)\ ]$$

where s is the item score. The relationship between constructs for RQ2 was tested with Spearman correlation (PU→BI and PEOU→BI) at the 5% level, and the open responses were analyzed thematically. Pairing TAM with SUS on a single system follows established practice in recent evaluations of interactive systems with expert users [27], and SUS reporting conventions and reference ranges follow the review of Vlachogianni and Tselios [28]. All TAM and SUS statistics are computed on a single basis, namely the responses that satisfy every inclusion criterion.

The user evaluation followed standard research-ethics practice for minimal-risk studies. Participation was voluntary and could be withdrawn at any point; respondents were informed of the study's purpose and of how their answers would be used, and gave consent before starting the questionnaire. No personally identifying information and no sensitive personal data were collected, no intervention was administered, and all results are reported only in aggregate. The ethical clearance covering the study is declared separately below.

III. RESULTS AND DISCUSSIONS

A. Ablation Results

Table 3 summarizes the eight configurations over the 40 test queries on the curated knowledge base. The first four columns range from 0 to 1 (higher is better), CVR is a percentage (lower is better), and the last column reports the Holm-adjusted p value of the Wilcoxon signed-rank test on the composite score against the complete configuration C8.

Table 3. Ablation results over 40 ground-truth queries.

Config	Context P@8	Rec P@5	Hit Rate	MRR	CVR (%)	p (vs C8)
C1 (baseline)	0.247	0.627	0.700	0.700	50.0	<0.001
C2 (MQR)	0.238	0.681	0.725	0.700	40.0	<0.001
C3 (MF)	1.000	1.000	1.000	1.000	0.0	<0.001
C4 (HyDE)	0.288	0.719	0.775	0.762	40.0	0.017
C5 (MQR+MF)	0.362	0.850	0.900	0.900	25.0	0.399
C6 (MQR+HyDE)	0.262	0.700	0.775	0.725	42.5	<0.001
C7 (MF+HyDE)	0.709	0.995	1.000	1.000	2.5	<0.001
C8 (complete)	0.366	0.878	0.925	0.912	20.0	reference

Source: authors (2026). Generator gemini-3.6-flash. p values are Holm-adjusted within the family of seven comparisons against C8.

Ranked by Context Precision@8, the order is C3 (1.000) > C7 (0.709) > C8 (0.366) > C5 (0.362) > C4 (0.288) > C6 (0.262) > C1 (0.247) > C2 (0.238). The complete configuration C8, hypothesized to be the strongest, lands mid-table, while the configuration that activates metadata filtering alone attains a perfect score on every precision metric and violates no stated constraint.

The first and most salient finding is the dominance of metadata filtering. All four MF-containing configurations (C3, C5, C7, C8) yield higher Context P@8 and lower CVR than the four without MF (C1, C2, C4, C6). Context P@8 spans 0.362 to 1.000 among MF configurations against 0.238 to 0.288 without MF, and CVR spans 0.0% to 25.0% against 40.0% to 50.0%. Filtering acts directly on structured attributes and therefore guarantees that every candidate satisfies the stated criteria, a guarantee semantic similarity alone cannot provide.

The second finding, contrary to expectation, is that adding query-expansion techniques on top of MF lowers precision. Among MF configurations, pure MF (C3, 1.000) beats MF+HyDE (C7, 0.709), which beats the complete pipeline (C8, 0.366) and MQR+MF (C5, 0.362). CVR rises along the same gradient, from 0.0% at C3 to 2.5% at C7, 20.0% at C8, and 25.0% at C5. Each added expansion path introduces candidates that never pass the metadata filter, and RRF then merges those unfiltered candidates back into the context.

The third finding is that MQR or HyDE alone, without MF, does not improve on the baseline. Relative to C1 (Context P@8 0.247; CVR 50.0%), pure MQR (C2) is marginally lower on precision (0.238) and pure HyDE (C4) marginally higher (0.288), and neither difference reaches significance once the comparison family is corrected. Expansion widens the candidate pool but does not make it more compliant.

The Wilcoxon test reinforces this reading. Against C8, C3 and C7 differ significantly with higher scores ($p < 0.001$), so their advantage over the complete pipeline is not sampling noise. C5 is the single configuration statistically indistinguishable from C8 ($p = 0.399$), which is consistent with both containing MQR and MF. Against the baseline C1, the pattern is cleaner still: the four MF-bearing configurations all differ significantly (C3 and C7 at $p < 0.001$, C5 at $p = 0.0003$, C8 at $p = 0.0002$), while none of the three without MF does (C2 $p = 0.760$, C4 $p = 0.381$, C6 $p = 0.760$).

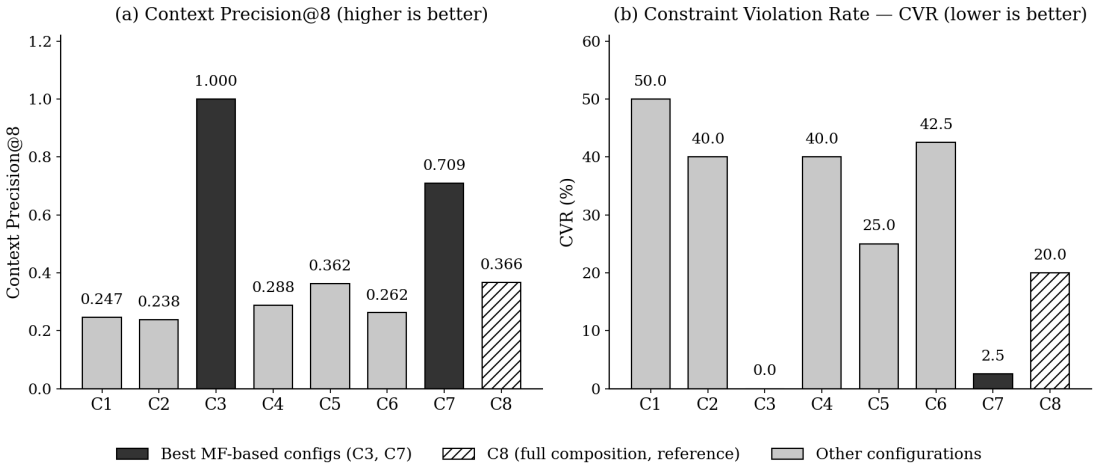


Fig. 4. Context Precision@8 and CVR across the eight configurations.

B. Cross-Model Replication

Table 4 places the primary run beside the full rerun on the second generator. The agreement is at the level of inference rather than of decimals. The six highest-ranked configurations hold exactly the same order in both models (C3 > C7 > C8 > C5 > C4 > C6); the only exchange is between C1 and C2 at the foot of the table, and that is precisely the pair that fails to differ significantly in either model. Most importantly, the significance pattern against the baseline is reproduced without exception: all four MF-bearing configurations outperform C1 significantly under both generators, and none of the three configurations without MF does under either. Four of four and zero of three, twice over.

Table 4. Cross-model replication over the same 40 ground-truth queries and knowledge base.

Config	Ctx P@8 Gemini	Ctx P@8 GPT	Comp. Gemini	Comp. GPT	p vs C1 Gemini	p vs C1 GPT
C1 (baseline)	0.247	0.247	0.555	0.542	reference	reference
C2 (MQR)	0.238	0.244	0.589	0.522	0.760	0.679
C3 (MF)	1.000	1.000	1.000	1.000	<0.001	<0.001
C4 (HyDE)	0.288	0.328	0.629	0.627	0.381	0.366
C5 (MQR+MF)	0.362	0.381	0.753	0.728	<0.001	<0.001
C6 (MQR+HyDE)	0.262	0.275	0.608	0.625	0.760	0.366
C7 (MF+HyDE)	0.709	0.722	0.936	0.931	<0.001	<0.001
C8 (complete)	0.366	0.394	0.776	0.791	<0.001	<0.001

Source: authors (2026). Generators gemini-3.6-flash and gpt-5.6-luna. p values are Holm-adjusted within the family of comparisons against the baseline C1. Composite is the unweighted mean of the five metrics with CVR entered as compliance.

The design's internal validity control behaves exactly as it must. Context Precision@8 and Hit Rate are identical across the two generators for C1 (0.247 and 0.700) and for C3 (1.000 and 1.000), and differ for all six configurations that call a language model on the retrieval path. C1 and C3 are the only two configurations whose retrieval is fully deterministic, so metrics computed on the retrieved context cannot vary with the generator, whereas MQR and HyDE construct their queries through the model and therefore must. What does shift within C1 is confined to Recommendation Precision@5, MRR, and CVR, which

are scored on the final answer rather than on the retrieved context. This confirms that the two runs differ only where they are designed to differ, and that the ranking is a property of the retrieval architecture rather than of a particular model.

C. Robustness to Query Form

Table 5 reports the same objective metrics on the numeral-free query set. Every configuration falls sharply. Context Precision@8 for C3 drops from 1.000 to 0.300 and its CVR rises from 0.0% to 72.5%, confirming that part of the filtering advantage in Table 3 is tied to explicitly constrained requests. The ordering, however, does not change: C3 remains significantly ahead of the baseline ($p = 0.016$) while adding MQR does not ($p = 0.493$). The prediction that query expansion would take over where filtering weakens is therefore not supported. The absolute values should be read against the difficulty of the set: on average only 6.5% of the corpus qualifies per query, so a random draw would score 0.065, making C1 1.9 times, C2 2.7 times, and C3 4.6 times better than chance.

Table 6 breaks the result down by group and explains the mechanism. Where a city or province survives in the text, filtering retains a wide margin: Context P@8 for C3 is 0.588 on implied-location queries and 0.575 on narrative queries, against 0.025 and 0.350 for the baseline. Where nothing can be extracted, the margin disappears: C3 scores 0.162, 0.125, and 0.089 on facility, budget, and size queries, which is level with or below the baseline, and it is exactly there that query expansion edges marginally ahead. Of the 40 requests, the parser recovered a location for 13, a star class for 1, and neither a capacity nor a budget for any. The advantage of structured filtering is thus proportional to how much of the constraint survives into the text, which is a testable statement rather than a caveat.

Table 5. Results on the 40 numeral-free natural-language queries.

Config	Context P@8	Rec P@5	Hit Rate	MRR	CVR (%)	p (vs C1)
C1 (baseline)	0.122	0.247	0.475	0.210	85.0	reference
C2 (MQR)	0.178	0.234	0.525	0.259	87.5	0.493
C3 (MF)	0.300	0.390	0.625	0.366	72.5	0.016

Source: authors (2026). Generator gemini-3.6-flash. p values from the Wilcoxon signed-rank test on the composite score against the baseline C1. CVR here counts violations of a rubric that interprets an unstated intent, so it is not interchangeable with CVR in Table 3.

Table 6. Context Precision@8 by query group on the numeral-free set.

Query group	n	C1	C2 (MQR)	C3 (MF)	Extractable constraint
A. Implied location	10	0.025	0.175	0.588	city recoverable from a colloquial name
B. Facility and atmosphere	10	0.162	0.188	0.162	none
C. Implied budget	8	0.078	0.078	0.125	none
D. Implied size	7	0.089	0.125	0.089	none
E. Narrative	5	0.350	0.400	0.575	province often named within the story

Source: authors (2026). Generator gemini-3.6-flash. The parser recovered a location for 13 of the 40 requests, a star class for 1, and neither a capacity nor a budget for any.

D. User Acceptance and Usability

Questionnaire data were collected online between 12 July and 20 August 2026, and 52 responses had been received at the cut-off date. Applying the inclusion and exclusion criteria removed 10 responses from outside the target population, two responses containing a blank questionnaire item, and two that failed the consistency check, leaving 38 valid responses, above the planned minimum of 30. Of the 42 responses from within the target population, 24 own an event-organizer business, 13 work as staff or freelancers, three are vendors or partners, one represents a company that commissions event organizers, and one is a tour guide; across all 52 responses, 36 (69.2%) report more than five years of professional experience.

The instrument met validity and reliability requirements. TAM showed very high reliability (Cronbach's alpha 0.964) with all ten items exceeding the critical value of the corrected item-total correlation, and SUS showed acceptable reliability (alpha 0.792) with nine of ten items valid; the single item below threshold is the reverse-worded complexity item, whose corrected correlation of 0.259 falls just under the critical 0.279. Reliability for SUS is computed after reverse-scoring the negatively worded items, as the scale requires. Table 7 reports the per-construct figures.

Table 7. Instrument validity and reliability.

Instrument / construct	Items	Cronbach α	Valid items
TAM, Perceived Usefulness	3	0.930	3/3
TAM, Perceived Ease of Use	3	0.874	3/3
TAM, Attitude	2	0.875	2/2
TAM, Behavioral Intention	2	0.911	2/2
TAM (overall)	10	0.964	10/10
SUS	10	0.792	9/10

Source: authors' primary data (2026). Alpha for SUS is computed after reverse-scoring items 2, 4, 6, 8, and 10. Critical values of the corrected item-total correlation at the 5% level: TAM (df = 50) = 0.273; SUS (df = 48) = 0.279.

TAM ratings over the 38 valid responses were high and consistent across constructs: perceived usefulness 4.46, perceived ease of use 4.47, attitude 4.47, and behavioral intention 4.41 on a five-point scale. Spearman correlations were positive and significant: perceived usefulness with behavioral intention at 0.837 and perceived ease of use with behavioral intention at 0.753, both at $p < 0.001$, so the acceptance relationships posited by TAM hold in this sample. The weaker of them, at 0.753, requires only 12 respondents to reach 80% power at the 5% level, so the 38 valid responses are comfortably sufficient for the association tests. Over the 38 valid responses the SUS mean was 77.7 with a median of 80.0, above the normative average of 68 [33]; 28 of the 38 respondents (74%) scored above that threshold and the 95% confidence interval, 72.7 to 82.7, lies entirely above it.

E. Discussion

Regarding RQ1, the effectiveness of Advanced RAG on constrained venue recommendation is determined chiefly by metadata filtering, not by stacking query-expansion techniques. MF acts as a hard-constraint gate that guarantees candidate admissibility, whereas MQR and HyDE widen semantic coverage at the cost of admissibility, and RRF then propagates that cost into the context. The finding is not an artifact of one model, since it survives a full rerun on an independently developed generator, and its scope is now explicit rather than assumed: it holds wherever constraints can be recovered from the request and weakens in proportion to how much of the constraint is left implicit. The pattern nuances query-expansion approaches that report gains from multi-query rewriting and fusion [9]: on open-ended generation those gains are plausible, but on constrained entity retrieval the same expansion dilutes a candidate set that filtering has already made correct. It also separates this work from a single-hotel RAG information service [29] and from tourism RAG systems that optimize semantic relevance without hard-constraint compliance [19], [20], [21], as well as from classical hotel recommenders that optimize rating prediction rather than constraint satisfaction [22]. What distinguishes the present study from all of these is not the application domain but the isolation of each component's contribution against verifiable ground truth.

Regarding RQ2 on the acceptance side, the uniformly high TAM scores, from 4.41 to 4.47 over 38 valid responses, support the premise that practitioners need a self-service venue-information channel, and a behavioral intention of 4.42 indicates readiness to adopt. The SUS mean of 77.7 sits above the normative average of 68 [26] and, with 38 valid respondents, its confidence interval excludes that threshold, so the system can be described as above average rather than merely not below it. The gap between very high TAM scores and a more moderate SUS score is itself informative: respondents value what the system delivers more than they are satisfied with how it is delivered. Thematic reading of the open responses locates the shortfall in interface presentation and process feedback rather than in recommendation quality, which is consistent with the ablation result that retrieval correctness is already governed by filtering.

IV. CONCLUSIONS

This study built Vento AI, an Advanced RAG chatbot for MICE venue recommendation over 1,741 Indonesian hotel venues in 36 provinces and 142 cities and regencies, and isolated the contribution of each retrieval technique through a full factorial ablation on 40 ground-truth queries. Answering RQ1, metadata filtering alone is the strongest configuration, reaching Context Precision@8 of 1.000 at a Constraint Violation Rate of 0%, while the complete pipeline that stacks all three techniques lands mid-table; the difference from the complete pipeline is significant under Holm-adjusted Wilcoxon testing. Two robustness checks bound and strengthen that answer. Rerunning the entire ablation on a second generator reproduced the inference exactly, with all four MF-bearing configurations significantly beating the baseline in both models and none of the three without MF doing so in either. A second query set written without numerals showed the filtering advantage shrinking to 0.300 while remaining significant, locating the boundary of the finding in how much of a constraint survives into the text of the request. Answering RQ2, evaluation with 38 valid respondents returned TAM means of 4.41 to 4.47 and a SUS mean of 77.7, above the normative average of 68. The practical implication is that a correctly placed structured filter contributes more to constrained recommendation quality than additional generative retrieval stages.

The study has limitations. The knowledge base and evaluation are scoped to Indonesian hotel venues, so the artifact's coverage is national, though the methodological finding is expected to transfer to other constrained entity-retrieval domains. Ground truth for the primary query set shares its constraint axes with the metadata filter, which is intrinsic to any exactly verifiable constrained-retrieval benchmark and is the reason the numeral-free query set was added; that second set, in turn, was run on three configurations and a single generator, so cross-model behavior on open-form requests remains untested. The constraint parser is rule-based and recognizes only explicitly written cities, provinces, regions, and star classes, so a model-based slot filler able to infer capacity and budget from vague phrasing is a natural next step. Finally, 365 venues in the curated knowledge base share a street address with a venue recorded in a different city and are flagged as pending field verification, and the user sample, while exceeding the planned minimum, remains purposive rather than probabilistic.

V. ACKNOWLEDGMENTS

The authors thank the event-industry practitioners who participated in the user evaluation. (Institutional and advisor acknowledgments to be added in the camera-ready version.)

VI. AUTHOR CONTRIBUTION

The first author contributed to the conceptualization of the study, construction and curation of the venue knowledge base, implementation of the retrieval-augmented generation pipeline, execution of the ablation experiments and user evaluation, and preparation of the original draft. The second author contributed to research supervision, methodology and validation, and critical review and editing of the manuscript. Both authors read and approved the final manuscript.

VII. ETHICAL CLEARANCE

The user-evaluation component collected opinions from adult professional respondents about a software prototype. Participation was voluntary and anonymous, informed consent was obtained before the questionnaire, no personally identifying or sensitive personal data were collected, no intervention was administered, and all findings are reported in aggregate. The study was conducted in accordance with the research-ethics rules of the authors' institution, and the ethical clearance reference is supplied to the editorial office in the submission metadata. The venue data underlying the knowledge base describe business entities rather than individuals and involve no human subjects.

VIII. DATA AVAILABILITY

The two query sets with their ground-truth labels and relevance rubrics, the ablation and statistical-analysis scripts for both generators, and the aggregated result tables reported in this article are openly deposited in a public research repository; the identifier is withheld during double-blind review and will be supplied in the camera-ready version. The venue knowledge base contains commercially sensitive negotiated rates and is available from the corresponding author on reasonable request.

IX. REFERENCES

[1] Y. K. Dwivedi, N. Pandey, W. Currie, and A. Micu, "Leveraging ChatGPT and other generative artificial intelligence (AI)-based applications in the hospitality and tourism industry: practices, challenges and research agenda," *International Journal of Contemporary Hospitality Management*, vol. 36, no. 1, pp. 1–12, Jan. 2024, doi: 10.1108/IJCHM-05-2023-0686.

[2] D. Hur, S. Lee, and H. Kim, "Are we ready for MICE 5.0? An investigation of technology use in the MICE industry using social media big data," *Tour. Manag. Perspect.*, vol. 43, 2022, doi: 10.1016/j.tmp.2022.100991.

[3] D. Hradecky, J. Kennell, W. Cai, and R. Davidson, "Organizational readiness to adopt artificial intelligence in the exhibition sector in Western Europe," *Int. J. Inf. Manage.*, vol. 65, 2022, doi: 10.1016/j.ijinfomgt.2022.102497.

[4] I. Martín-Rojo and A. I. Gaspar-González, "The impact of social changes on MICE tourism management in the age of digitalization: A bibliometric review," *Review of Managerial Science*, vol. 19, no. 7, pp. 2211–2234, 2025, doi: 10.1007/s11846-024-00751-3.

[5] S. Klinmalai and P. Kaewlai, "Assessment and selection criteria for MICE destinations and convention centers in Thailand," *Journal of Architectural/Planning Research and Studies*, vol. 21, no. 2, pp. 245–262, 2024, doi: 10.56261/jars.v21.258316.

[6] E. P. A. P. F. P. V. K. N. G. H. K. M. L. W. Y. T. R. S. R. D. K. Patrick Lewis, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, 2020, pp. 9459–9474.

[7] Y. X. X. G. K. J. J. P. Y. B. Y. D. J. S. M. W. and H. W. Yunfan Gaoa, "Retrieval-Augmented Generation for Large Language Models: A Survey".

[8] W. F. et al., "A survey on RAG meeting LLMs: Towards retrieval-augmented large language models," in *Proc. 30th ACM SIGKDD*, 2024. doi: 10.1145/3637528.3671470.

[9] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, "Query rewriting for retrieval-augmented large language models," in *Proc. 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 5303–5315. doi: 10.18653/v1/2023.emnlp-main.322.

[10] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," in *Proc. 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. doi: 10.18653/v1/2023.acl-long.99.

[11] K. A. F. et al., "Hybrid recommender system for tourism based on big data and AI: A conceptual framework," *Big Data Mining and Analytics*, vol. 4, no. 1, pp. 47–55, 2021, doi: 10.26599/BDMA.2020.9020015.

[12] G. V Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in *Proc. 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009, pp. 758–759. doi: 10.1145/1571941.1572114.

[13] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 150–158. doi: 10.18653/v1/2024.eacl-demo.16.

[14] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of retrieval-augmented generation: A survey," in *Big Data (BigData 2024), Communications in Computer and Information Science, vol. 2301*, Singapore: Springer, 2025, pp. 102–120. doi: 10.1007/978-981-96-1024-2_8.

[15] C. Wei, K. Duan, S. Zhuo, H. Wang, S. Huang, and J. Liu, "Enhanced Recommendation Systems with Retrieval-Augmented Large Language Model (Abstract Reprint)," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 47, pp. 39904–39904, Mar. 2026, doi: 10.1609/aaai.v40i47.41419.

[16] J. L. et al., "How can recommender systems benefit from large language models: A survey,"

ACM Trans. Inf. Syst., vol. 43, no. 2, 2025, doi: 10.1145/3678004.

[17] Z. Zhao *et al.*, "Recommender Systems in the Era of Large Language Models (LLMs)," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 11, pp. 6889–6907, Nov. 2024, doi: 10.1109/TKDE.2024.3392335.

[18] C. Wei, K. Duan, S. Zhuo, H. Wang, S. Huang, and J. Liu, "Enhanced recommendation systems with retrieval-augmented large language model," *Journal of Artificial Intelligence Research*, vol. 82, pp. 1147–1173, 2025, doi: 10.1613/jair.1.17809.

[19] R. Karlović, M. Rovis, A. Smajić, L. Sever, and I. Lorencin, "Context-aware tourism recommendations using retrieval-augmented large language models and semantic re-ranking," *Electronics (Basel).*, vol. 14, no. 22, 2025, doi: 10.3390/electronics14224448.

[20] A. Banerjee, A. Satish, and W. Wörndl, "Enhancing Tourism Recommender Systems for Sustainable City Trips Using Retrieval-Augmented Generation," 2025, pp. 19–34. doi: 10.1007/978-3-031-87654-7_3.

[21] S. Song, C. Yang, L. Xu, H. Shang, Z. Li, and Y. Chang, "TravelRAG: A tourist attraction retrieval framework based on multi-layer knowledge graph," *ISPRS Int. J. Geoinf.*, vol. 13, no. 11, 2024, doi: 10.3390/ijgi13110414.

[22] Q. Y. Shambour, A. A. Abu-Shareha, and M. M. Abualhaj, "A hotel recommender system based on multi-criteria collaborative filtering," *Information Technology and Control*, vol. 51, no. 2, pp. 390–402, 2022, doi: 10.5755/j01.itc.51.2.30701.

[23] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[24] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989, doi: 10.2307/249008.

[25] A. M. Deshmukh and R. Chalmeta, "Validation of System Usability Scale as a usability metric to evaluate voice user interfaces," *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/peerj-cs.1918.

[26] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the System Usability Scale," *Int. J. Hum. Comput. Interact.*, vol. 24, no. 6, pp. 574–594, 2008, doi: 10.1080/10447310802205776.

[27] A. Lekova, "A design-based research approach to streamline the integration of high-tech assistive technologies in speech and language therapy," *Technologies (Basel).*, vol. 13, no. 7, 2025, doi: 10.3390/technologies13070306.

[28] P. Vlachogianni and N. Tselios, "Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review," *Journal of Research on Technology in Education*, vol. 54, no. 3, pp. 392–409, 2022, doi: 10.1080/15391523.2020.1867938.

[29] M. H. Wijaya and H. Jayadianti, "Implementation of RAG-based LLM chatbot for hotel services: A case study," *International Journal of Multidisciplinary Sciences and Arts*, vol. 5, no. 1, pp. 275–281, 2026, doi: 10.47709/ijmdsa.v5i1.7927.