

Adaptive hybrid AI-based retrieval-augmented generation for construction and seismic regulatory question answering

D. A. Bekmirzaev¹, R. R. Yuldoshev², E. A. Kosimov³, A. M. Ismoilov⁴, S. S. Shaumarov⁵

¹Institute of Mechanics and Seismic Stability of Structures named after M.T. Urazbaev, Tashkent, Uzbekistan

^{2, 3, 4}National Research Institute for Seismic Safety and Resilient Construction under Tashkent Institute of Architecture and Civil Engineering, Tashkent, Uzbekistan

⁵Tashkent State Transport University, Tashkent, Uzbekistan

¹Corresponding author

E-mail: ¹diyorbek_84@mail.ru, ²aloqa@suvoriy.com, ³elbekqosimov1985@gmail.com,

⁴aismoilov802@gmail.com, ⁵shoumarovss@gmail.com

Received 21 April 2026; accepted 18 August 2026; published online 30 September 2026

DOI <https://doi.org/10.21595/mme.2026.26583>



Copyright © 2026 D. A. Bekmirzaev, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract. Construction and seismic regulatory documents, such as the Uzbek QMQ (Qurilish Me'yorlari va Qoidalari – Construction Norms and Rules) and ShNQ (Shaharsozlik Normalari va Qoidalari – Urban Planning Norms and Rules), contain structured clause identifiers, numerical constraints and domain-specific terminology that conventional dense retrieval methods often fail to match reliably, while retrieval-augmented generation (RAG) systems additionally suffer from retrieval latency that limits interactive use. This paper proposes an Adaptive Hybrid RAG Model (AHRM) that combines semantic (Sentence-BERT with FAISS), keyword-based (BM25) and rule-based retrieval within a non-linear fusion function whose weights are assigned adaptively according to the detected query type, and deploys it as a real-time question-answering chatbot for Uzbek construction and seismic regulations. On a corpus of 120 regulatory documents (approximately 85,000 text chunks), AHRM reaches a Precision@3 of 0.88, an MRR of 0.85 and an answer accuracy of 91 %, outperforming dense-only, keyword-only and statically weighted hybrid baselines, with a mean index-search latency of 0.6 ms per query over the filtered candidate pool, measured separately from query embedding and answer generation. The results indicate that combining rule-aware retrieval with query-adaptive weighting improves the reliability of regulatory question answering in safety-critical engineering domains.

Keywords: retrieval-augmented generation, hybrid retrieval, semantic search, FAISS, construction standards, seismic regulations, question answering, adaptive weighting.

1. Introduction

Retrieval-Augmented Generation (RAG) is a framework that combines document retrieval with large language model generation in order to produce responses grounded in external knowledge sources [1]. Instead of relying solely on parametric knowledge stored in model weights, RAG first retrieves relevant document chunks from a corpus and then uses these chunks as contextual evidence for answer generation. This paradigm has been shown to reduce hallucinations, improve factual correctness, and enhance interpretability in question-answering systems [2, 3].

RAG has been successfully applied in open-domain question answering [2, 4] and legal document analysis [5, 6], while hybrid lexical-semantic retrieval models improve first-stage retrieval by combining exact-match and embedding signals [7]. Its effectiveness is particularly important in structured knowledge domains, where answers must be supported by precise references rather than approximate semantic similarity alone. In such settings, retrieval quality directly determines the reliability of the generated response.

This challenge is especially significant in construction and seismic regulation, where accurate retrieval of standards, clauses, and engineering requirements is critical; even minor inaccuracies may lead to unsafe engineering decisions or regulatory violations. In particular, construction and seismic regulatory documents such as QMQ (Qurilish Me'yorlari va Qoidalari – Construction Norms and Rules), which govern structural design and construction safety, and ShNQ (Shaharsozlik Normalari va Qoidalari – Urban Planning Norms and Rules), which define urban planning and seismic safety requirements, contain highly structured numerical identifiers, technical constraints, and domain-specific terminology. These characteristics make retrieval particularly challenging for standard dense retrieval methods, which may fail to capture exact clause references, code numbers, or formally defined engineering expressions.

The practical stakes are high in a seismically active country such as Uzbekistan, where these normative documents govern the entire life cycle of transport infrastructure: the estimation of the seismic response of monolithic overpasses [8], the influence of environmental factors such as seasonal temperature on continuous structures [9], the design resistance of soils and the behaviour of reinforced-soil bridge abutments [10, 11], the strengthening of existing shallow and pile foundations [12], the seismic performance assessment of ordinary building stock in seismic zones [13], the seismodynamic response of segmented buried lifelines under recorded earthquake excitations [14], and the vulnerability analysis and instrumented monitoring of in-service bridges [15-17], together with the metrological calibration of the accelerometers on which such monitoring depends [18], all rely on rapid and precise reference to the applicable QMQ and ShNQ provisions.

More broadly, the volume and pace of the technical literature that must be read alongside such norms continue to grow – from comprehensive reviews of structural analysis theories [19] and studies of the dynamic behaviour of composite structures [20] to hybrid AI-based techniques for damage prediction and forecasting in engineering structures [21, 22], risk-oriented predictive diagnostics built on vibration signal analysis [23], computer-vision frameworks for real-time incident recognition in transport systems [24] and satellite-based geospatial monitoring of extended infrastructure [25] – which further motivates automated, precise retrieval over large document collections.

Although the retrieval techniques discussed above originate in general-purpose and open-domain settings, the design presented in this paper is deliberately specialised for construction and seismic regulatory content, for four reasons. First, such documents are organised around rigid clause identifiers (e.g., QMQ 2.01.03-96, section, table and appendix numbers), which generic dense retrieval treats as ordinary tokens but which the rule-based channel exploits as first-class evidence. Second, the domain is safety-critical: retrieving an adjacent but wrong clause can translate directly into an incorrect engineering decision, which motivates the emphasis on precision at small k and the refusal mechanism described in Section 4. Third, the corpus is inherently bilingual (Uzbek and Russian), which dictates the choice of a multilingual embedding model. Fourth, regulatory answers must remain verifiable against numbered provisions, which the clause-citing generation procedure guarantees. The architecture itself, however, is not tied to Uzbek regulations; its transferability to other countries, datasets and application areas is discussed in Section 6.

Despite these advantages, two obstacles currently limit the deployment of RAG in structured domains such as engineering regulation.

The first obstacle is latency. Dense vector representations and transformer-based models are computationally expensive. Querying large document corpora with high-dimensional embeddings can result in significant delays, making real-time interaction difficult [26, 27]. This is particularly critical for interactive systems, such as engineering regulatory systems or online advisory tools, where users expect near-instantaneous responses.

The second obstacle is uncertainty. RAG systems can generate incorrect or misleading answers when retrieved documents are irrelevant or insufficiently precise [1, 4]. In structured domains such as law, failing to retrieve exact references may result in engineering risks. Purely dense retrieval

approaches may also fail to capture exact keyword matches, technical terminology, or domain-specific expressions, thereby increasing uncertainty in generated responses [1].

To address these obstacles, hybrid retrieval methods have been proposed that combine complementary evidence sources: semantic embeddings, which capture contextual meaning using models such as BERT or Sentence-BERT [26, 28]; keyword-based retrieval, which guarantees exact term matching for domain-specific expressions [29]; and rule-based patterns, which capture structured references such as article numbers and clause identifiers. By prioritising relevant chunks and reducing the effective search space, such combinations can mitigate both latency and uncertainty [7].

Building on these observations, this paper introduces the Adaptive Hybrid RAG model for structured construction and seismic regulatory domains. The key contributions are as follows:

- 1) A multi-stage retrieval pipeline that sequentially applies rule-based filtering, keyword overlap detection, and semantic similarity search.
- 2) An adaptive weighting mechanism that dynamically adjusts the influence of semantic, keyword, and rule-based retrieval depending on the query type.
- 3) Integration with a real-time ReactJS-based chatbot, demonstrating practical applicability in construction and seismic regulatory advisory tasks.
- 4) Empirical results showing reduced response latency (~ 1 ms per query for the index-search stage) and improved answer accuracy compared to baseline RAG implementations (Section 6).

2. Related work

FAISS (Facebook AI Similarity Search) is a library optimized for large-scale vector similarity search [30]. It supports approximate nearest neighbor (ANN) search using techniques such as inverted file (IVF) indexing and product quantization, markedly reducing search time while maintaining retrieval accuracy [31].

Experimental setup: Experiments were conducted on a workstation with an Intel Core i7 CPU, 16 GB RAM, running FAISS IndexFlatIP. In the deployed sequential configuration (Section 3.5), this exact inner-product scoring is applied to the pre-filtered candidate subset rather than to the full index on every query – which is what the reported latency reflects – so the exhaustive property of IndexFlatIP holds over the candidate pool presented to the semantic stage.

In our experiments on structured construction and seismic regulatory documents (QMQ and ShNQ), FAISS enabled retrieval from the indexed regulatory chunks well under a millisecond per query.

The reported latency corresponds to the average index-search time per query, excluding query embedding, for top-10 nearest neighbors on an Intel Core i7 CPU.

The experiment was conducted using Sentence-BERT embeddings with dimensionality $d = 384$ and a FAISS IndexFlatIP configuration.

RAG integrates retrieval and generation to produce context-aware answers [1]. Retrieved document chunks condition the generative model, reducing hallucinations and improving factual correctness [2].

However, several challenges remain:

- Latency: Dense retrieval is computationally expensive, particularly for large-scale regulatory corpora [32].
- Uncertainty: Generative models may produce incorrect references if retrieved chunks are irrelevant or incomplete [33].

Compared to prior work, the proposed system offers the following contributions:

- Dynamic adjustment of semantic and keyword weighting based on query characteristics.
- Demonstration of real-time performance within a ReactJS-based construction and seismic regulatory environment.
- Empirical improvements in retrieval precision and latency over dense, keyword-only and statically weighted baselines, quantified in Section 6.

A further line of recent work employs large language models not only for answer generation but inside the retrieval loop itself: LLM-derived dense representations, LLM-based re-ranking of candidate lists [34], fully generative retrieval in which the model directly produces document identifiers, and agentic, iterative retrieval schemes; a systematic taxonomy of these naive, advanced and modular RAG designs is given in [35]. These approaches can be effective, but they make every retrieval decision as computationally expensive and as opaque as a forward pass of the language model, which is difficult to justify in a safety-critical regulatory setting that requires auditable decisions, low latency and fully local deployment. The approach proposed here deliberately takes the opposite route: retrieval is performed by three inexpensive, deterministic and mathematically analysable channels (Section 3), and the large language model – DeepSeek-R1 [36] served locally via Ollama – is confined to the final, well-delimited role of composing an answer from already-retrieved, verifiable evidence [1, 2, 37]. The novelty of the contribution therefore lies not in scaling the generative component, but in the query-adaptive, rule-aware organisation of the evidence that reaches it. In the deployed configuration, generation is fully specified by components already described: the locally served model receives the deterministic prompt of Section 3.5, is constrained to the retrieved context, must cite clause identifiers, and is bypassed by an explicit refusal whenever the maximum fused score falls below the confidence threshold – so that the retrieval side and the generation-side interface – prompt construction, context restriction, citation requirement and refusal – are fully documented design decisions; the internal generation process of the LLM itself remains, as for any large language model, stochastic rather than transparent, which is precisely why it is confined to evidence already retrieved and verified. The reliance on a locally served open-weight model, rather than a hosted API, likewise follows directly from the confidentiality constraint on the regulatory corpus stated in Section 4.

3. Proposed method

This section presents the proposed Adaptive Hybrid Retrieval-Augmented Generation Model (AHRM), which integrates semantic, keyword-based, and rule-based retrieval strategies into a unified and dynamically adaptive framework. Unlike traditional RAG systems that rely solely on dense vector retrieval [1], the proposed method introduces a multi-stage retrieval pipeline with query-adaptive weighting and low-latency retrieval capabilities.

3.1. System overview

The overall architecture consists of the following stages:

- 1) Document Processing (PDF extraction and cleaning).
- 2) Text Chunking and Representation.
- 3) Vector Indexing (FAISS).
- 4) Hybrid Retrieval Engine.
- 5) Adaptive Weighting Mechanism.
- 6) Response Generation (LLM).

Let the document corpus be defined as:

$$D = \{d_1, d_2, \dots, d_n\}. \quad (1)$$

Each document is split into smaller chunks:

$$C_i = \{c_{i1}, c_{i2}, \dots, c_{im}\}, \quad (2)$$

$$c_i \rightarrow v_i \in R^d. \quad (3)$$

Similarly, a query q is mapped into the same vector space:

$$q \rightarrow v_q \in R^d . \quad (4)$$

3.2. Retrieval channels: semantic, keyword-based and rule-based scoring

Semantic similarity is computed using cosine similarity, defined in Eq. (5):

$$S_{semantic}(q, c_i) = \frac{v_q \cdot v_i}{\|v_q\| \|v_i\|}. \quad (5)$$

After L2-normalization, cosine similarity reduces to:

$$S_{semantic}(q, c_i) = v_q \cdot v_i. \quad (6)$$

For the IndexFlatIP configuration, retrieval complexity is $O(n)$.

To improve lexical matching, we adopt the BM25 scoring function [29], given in Eq. (7):

$$S_{keyword}(q, c_i) = \sum_{t \in q} IDF(t) \cdot \frac{f(t, c_i)(k_1 + 1)}{f(t, c_i) + k_1 \left(1 - b + b \frac{|c_i|}{avgdl}\right)}, \quad (7)$$

where $f(t, c_i)$ is the term frequency, $|c_i|$ is chunk length, and $avgdl$ is the average document length.

For structured engineering domains such as construction and seismic regulations (QMQ and ShNQ), we define a weighted rule-based scoring function, Eq. (8):

$$S_{rule}(q, c_i) = \sum_{p \in P_q} w_p \cdot match(p, c_i), \quad (8)$$

where pattern weights are defined by Eq. (9) as:

$$w_p = \frac{freq(p)}{\sum_{p' \in P} freq(p')}. \quad (9)$$

The matching function is given by Eq. (10) as:

$$match(p, c_i) = \begin{cases} 1, & \text{full match,} \\ \delta, & \text{partial match,} \\ 0, & \text{no match,} \end{cases} \quad (10)$$

where $\delta \in (0,1)$ represents partial matching weight.

The pattern set P comprises regular expressions reflecting the clause-numbering conventions of Uzbek regulatory documents: full document codes (e.g., QMQ 2.01.03-96), section and clause references, table and appendix identifiers, and numeric parameter expressions such as seismic intensity values. Structured queries containing such identifiers are precisely those for which the rule-based channel contributes most strongly; its quantitative impact on retrieval accuracy is isolated in the ablation study (Section 5.3).

3.3. Adaptive fusion and query-adaptive weighting

We define a non-linear scoring function to combine multiple retrieval signals, Eq. (11):

$$Score(q, c_i) = \sigma(w_1 S_{semantic} + w_2 S_{keyword} + w_3 S_{rule} + w_4 S_{semantic} \cdot S_{rule}), \quad (11)$$

where $\sigma(\cdot)$ is the sigmoid activation function.

The relative influence of the three retrieval channels is assigned per query as:

$$(w_1, w_2, w_3) = \text{softmax}(W \cdot \phi(q)), \quad (12)$$

where $\phi(q)$ is a vector of binary indicators obtained by thresholding simple query statistics – the presence of a document-code or clause pattern, an indicator of high digit density, and an indicator of long queries – with the indicator thresholds recorded in the released configuration, and W is a fixed, expert-defined mapping matrix that raises the rule-based weight for structured queries and the semantic weight for natural-language queries. In the current implementation W is not trained; learning it from annotated query logs is identified as future work (Section 7). The interaction coefficient $w_4 \geq 0$ in Eq. (11) is a fixed, non-negative configuration constant chosen together with the mapping W ; it is not produced by the softmax of Eq. (12) and is therefore treated as a separate bounded parameter in the analysis of Section 3.4.

3.4. Analytical properties of the fusion model

The fused score of Eq. (11) possesses several properties that make its behaviour predictable and analysable. Because the outer function is the logistic sigmoid $\sigma(z) = \frac{1}{1+e^{-z}}$, the fused score is bounded, $\text{Score} \in (0,1)$, for any values of the channel scores, so that no single channel can produce an unbounded contribution to the ranking. Differentiating Eq. (11) with respect to each channel score yields the partial derivatives given in Eq. (13):

$$\begin{aligned} \frac{\partial \text{Score}}{\partial S_{\text{semantic}}} &= \sigma'(z) \cdot (w_1 + w_4 \cdot S_{\text{rule}}), \\ \frac{\partial \text{Score}}{\partial S_{\text{keyword}}} &= \sigma'(z) \cdot w_2, \\ \frac{\partial \text{Score}}{\partial S_{\text{rule}}} &= \sigma'(z) \cdot (w_3 + w_4 \cdot S_{\text{semantic}}). \end{aligned} \quad (13)$$

The bilinear interaction term $w_4 \cdot S_{\text{semantic}} \cdot S_{\text{rule}}$ acts as a conjunctive amplifier: its contribution is significant only when the semantic and the rule-based channels return high scores simultaneously, and it vanishes whenever either factor is close to zero. For structured queries this rewards chunks that are both topically close to the query and contain the exact clause identifier – precisely the behaviour required in regulatory retrieval – whereas for natural-language queries, where $S_{\text{rule}} \approx 0$, the model degrades gracefully to a weighted combination of semantic and lexical evidence.

The query-adaptive weights of Eq. (12) are produced by the softmax operator and therefore satisfy $w_i > 0$ and $\sum w_i = 1$ for every query, which fixes the relative budget of the three linear terms. The channel scores themselves, however, live on different scales – a signed cosine, a non-negative BM25 score bounded by B_q and a non-negative rule score bounded by $W_p(q)$ – so the pre-activation is not a convex combination of commensurate quantities; what the analysis below relies on is the weaker but sufficient property that it is bounded for every query, as summarised in Eq. (14).

The keyword channel inherits the saturation behaviour of BM25: for a fixed document length, the contribution of a term grows with its frequency but is bounded by the limit $(k_1 + 1) \cdot \text{IDF}(t)$ as the frequency tends to infinity, so that k_1 controls how quickly repeated occurrences of a term stop adding evidence, while $b \in [0,1]$ interpolates between no document-length normalisation ($b = 0$) and full normalisation ($b = 1$). These are exactly the two parameters, together with the partial-match weight δ of Eq. (10), whose ranges are identified in Section 5.3.

Two further properties are worth stating. First, because the sigmoid is strictly increasing, the

ranking of candidates induced by the fused score coincides exactly with the ranking induced by the pre-activation z itself; the non-linearity therefore does not alter the retrieval order but maps the score into $(0, 1)$, the scale on which the refusal mechanism of Section 4 operates. Because the bound of Eq. (14) is query-dependent (through B_q and $W_P(q)$), large lexical scores can drive the sigmoid close to saturation; the numerical value of the refusal threshold is therefore a configuration parameter chosen on a development set rather than a universal constant. The roles remain cleanly separated: the linear-plus-interaction form of z determines which chunks are returned, while σ provides the common $(0, 1)$ scale on which the answering decision is made.

Second, all three channel scores are bounded for every fixed query, as summarised in Eq. (14): with the L_2 -normalised embeddings of Eq. (6) the semantic score is a cosine similarity in $[-1, 1]$; the BM25 score is bounded by the finite query-dependent constant B_q ; and the rule score is bounded by the total weight $W_P(q)$ of the patterns matched in the query, so that the pre-activation z is finite for every query. Small perturbations of the evidence therefore produce proportionally small changes of the score: since $\sigma'(z) \leq \frac{1}{4}$, Eq. (13) implies that the fused score is Lipschitz-continuous in each channel, with the constants $\frac{w_2}{4}$ for the keyword channel, $\frac{w_3+w_4}{4}$ for the rule channel and $\frac{w_1+w_4 \cdot W_P(q)}{4}$ for the semantic channel:

$$\begin{aligned} -1 &\leq S_{\text{semantic}} \leq 1, \\ |S_{\text{keyword}}| &\leq B_q := \sum_{t \in q} (k_1 + 1) \cdot |IDF(t)|, \\ 0 &\leq S_{\text{rule}} \leq W_P(q), \\ |z| &\leq w_1 + w_2 \cdot B_q + (w_3 + w_4) \cdot W_P(q). \end{aligned} \tag{14}$$

The keyword bound is defined through absolute IDF values, so that it holds for any IDF variant – including formulations in which the IDF of very frequent terms may become negative – while the rule bound equals the total weight of the patterns matched in the query.

Finally, the computational cost of the fusion stage is negligible compared with candidate generation. Exact inner-product search over the n candidate vectors presented to the search stage, each of dimension d , costs $O(n \cdot d)$ per query (Section 4); BM25 scoring is linear in the total number of postings of the query terms; and rule matching is linear in the number of patterns and the query length. The fusion itself is $O(1)$ per candidate, so the overall per-query complexity is dominated by the index search, which is consistent with the latency figures reported in Section 6.

These observations can be summarised as six properties: (P1) the fused score is bounded in $(0, 1)$; (P2) it is strictly monotone in the semantic and keyword channels, and in the rule channel under the explicit sign condition stated above; (P3) the interaction term implements a conjunction of semantic and rule evidence; (P4) the softmax normalisation fixes the budget of the linear terms ($w_1 + w_2 + w_3 = 1$), with the fixed interaction coefficient w_4 treated as a separate bounded parameter; (P5) the candidate ranking is invariant to the sigmoid and is decided by the pre-activation alone; and (P6) the score is Lipschitz-continuous with explicit per-channel constants. Together, P1-P6 characterise the fusion model as a bounded, stable and order-preserving scoring rule whose behaviour can be predicted analytically for any query type – a level of transparency that purely learned, end-to-end retrieval scorers do not generally provide.

Proposition 1 (properties of the fused score). For every query q and candidate chunk c : (i) $\text{Score} \in (0, 1)$; (ii) Score is strictly increasing in S_{semantic} and S_{keyword} ; it is strictly increasing in S_{rule} whenever $w_3 + w_4 \cdot S_{\text{semantic}} > 0$ — in particular whenever $S_{\text{semantic}} \geq 0$, for every $S_{\text{semantic}} \in [-1, 1]$ when $w_4 < w_3$, and for all cases when $w_4 = 0$; (iii) the ranking of candidates induced by Score coincides with the ranking induced by the pre-activation z ; (iv) the pre-activation is bounded for every query as stated in Eq. (14); and (v) Score is Lipschitz-continuous in each channel with the explicit constants stated above. Proof: (i) follows from $\sigma \in (0, 1)$; (ii) from

$\sigma'(z) > 0$, the derivatives of Eq. (13), $S_{rule} \geq 0$ and the stated sign condition for the rule channel; (iii) from the strict monotonicity of σ ; (iv) from the channel bounds $S_{semantic} \in [-1, 1]$, $S_{keyword} \leq B_q$ and $S_{rule} \leq W_P(q)$; and (v) from $\sigma'(z) \leq \frac{1}{4}$ applied to the derivatives of Eq. (13).

□

3.5. Retrieval pipeline and answer generation

The proposed multi-stage retrieval pipeline operates as follows:

- 1) Rule-based pre-filtering using S_{rule} .
- 2) BM25-based keyword filtering using $S_{keyword}$.
- 3) FAISS-based semantic retrieval using $S_{semantic}$.

This hierarchical filtering reduces the effective search space and improves practical retrieval efficiency. These numbered stages describe the order of computation: the earlier, inexpensive filters successively narrow the candidate pool, and the three channel scores entering the fusion of Eq. (11) are then evaluated for the candidates that survive filtering. Fig. 1 therefore depicts the logical scoring relationship among the channels, while the list above describes their sequential execution; the two views are complementary descriptions of the same pipeline rather than alternative algorithms. A stage acts as a filter only when it fires: if no rule pattern matches the query, the rule stage leaves the candidate set unchanged and simply contributes a zero rule score to Eq. (11), and the keyword stage behaves in the same way for query terms absent from the index, so the pipeline degrades gracefully to the remaining channels.

Top- k retrieved chunks are used as context, Eq. (15):

$$Context = \{c_{i1}, c_{i2}, \dots, c_{ik}\}. \quad (15)$$

The final response is generated using a large language model, Eq. (16):

$$Answer = LLM(q, Context). \quad (16)$$

This follows the RAG paradigm while enhancing retrieval robustness and efficiency [1, 2].

The augmentation step itself is performed deterministically. The prompt supplied to the generative model is assembled from three components: a fixed system instruction that restricts the model to the supplied context and requires clause identifiers to be cited; the top- k retrieved chunks, ordered by decreasing fused score and prefixed with their document and clause identifiers; and the user query. If the concatenated context exceeds the token budget of the model, the lowest-ranked chunks are truncated first, so that the strongest evidence is always preserved. Augmentation therefore modifies neither the model weights nor its decoding procedure; it only conditions generation on retrieved, verifiable evidence, which is what distinguishes the augmented configuration analysed in Section 6 from a non-augmented (closed-book) use of the same model.

3.6. System architecture

This section describes the overall architecture of the proposed Adaptive Hybrid Retrieval-Augmented generation model (AHRM), highlighting the end-to-end data flow from PDF document ingestion to real-time API responses. The overall architecture is summarised in Fig. 1.

The indexing pipeline instantiates the formalism of Section 3 directly: the corpus and chunk structure follow Eqs. (1-2); chunks and queries are embedded into the shared vector space of Eqs. (3-4) using the multilingual Sentence-BERT model [28]; and L2-normalisation of the stored vectors reduces the cosine similarity of Eq. (5) to the inner product of Eq. (6), which is exactly the quantity computed by the FAISS index [30]. No additional formalism is therefore required at the implementation level.

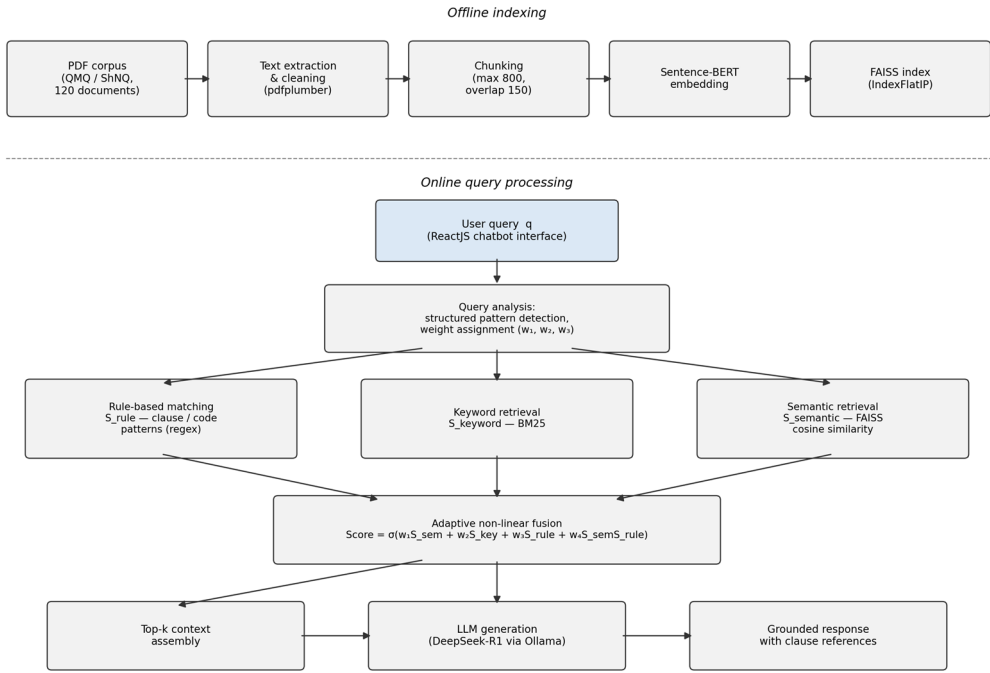


Fig. 1. Overall architecture of the proposed AHRM: offline indexing of QMQ/ShNQ documents and online multi-channel retrieval with adaptive fusion and LLM-based answer generation

Chunking preserves semantic integrity and table/formula context. In our implementation, a hybrid text splitter ensures that paragraphs, tables, and formulas remain within single chunks, preventing context fragmentation.

Normalized vectors are stored in a FAISS IndexFlatIP index for fast nearest neighbor search [30, 31]. The IndexFlatIP configuration performs exact (exhaustive) inner-product search, whose cost grows linearly with the number of stored vectors; at the current corpus scale (approximately 85,000 chunks) this remains fast in practice, while for substantially larger corpora approximate structures such as IVF or HNSW [30, 31] provide sub-linear scaling at a controlled loss of accuracy.

The proposed retrieval framework integrates semantic similarity, keyword-based matching, and rule-based filtering within a unified pipeline. Instead of combining these signals using a linear formulation, the model employs a non-linear scoring function (see Section 3.3) to capture interactions between semantic and structured features.

This design is particularly effective for structured engineering queries, where numerical identifiers (e.g., QMQ clause numbers), seismic parameters, and domain-specific terminology must be accurately matched. This approach improves retrieval robustness in safety-critical domains, where precise matching of regulatory constraints directly impacts decision reliability.

The top- k retrieved chunks are served through a RESTful API, which is integrated with a ReactJS-based chatbot interface:

1. The user submits a query through the chatbot interface.
2. The API performs efficient retrieval and scoring of candidate chunks.
3. The top- k ranked chunks are concatenated to form the context.
4. The constructed context is passed to the LLM (DeepSeek-R1 via Ollama) to generate the final response.

The DeepSeek-R1 model served via Ollama was selected for three practical reasons: it is an open-weight model that can be deployed entirely on local infrastructure, which is a strict requirement because regulatory documents cannot be transmitted to external cloud services; its

reasoning-oriented architecture is well suited to multi-step regulatory queries; and it makes Uzbek- and Russian-language input feasible; a systematic assessment of its behaviour on domain-specific terminology is part of the planned extended evaluation (Section 7). A systematic comparison with alternative generative models is beyond the scope of this study and is identified as future work (Section 7).

To reduce the risk of hallucinated answers when retrieved chunks are only partially relevant, three safeguards are applied. First, the generation prompt explicitly instructs the model to answer strictly on the basis of the supplied context. Second, queries whose maximum fused retrieval score falls below a confidence threshold are answered with an explicit refusal rather than a speculative response. Third, generated answers cite the identifiers of the source clauses, allowing the user to verify each statement against the original document.

Since the chatbot operates as a networked, interactive client-server application, its deployment also takes into account security considerations discussed for interactive communication media algorithms in networking applications [38] and for secure data transmission in sensor-based IoT deployments [39].

Because the underlying corpus combines documents in Uzbek and Russian and user queries may mix both languages, a multilingual Sentence-BERT model is used for embedding, so that queries and chunks in different languages are mapped into a shared vector space; the keyword and rule-based channels operate on the surface forms present in the documents. A systematic evaluation of cross-lingual retrieval quality is left for future work (Section 7).



Fig. 2. Real-time chatbot interface of the proposed AHRM system for construction and seismic regulatory question answering. The interface illustrates user query input, retrieval of relevant QMQ and ShNQ clauses, and generation of structured, context-aware responses

In our measurements, the index-search stage (candidate scoring and ranking across the FAISS, BM25 and rule channels) required approximately 0.6 ms per query on average over the filtered candidate pool produced by the pre-filtering stages of Section 3.5 (not an exhaustive scan of the full index) (top- k results, Intel Core i7 CPU). This figure refers exclusively to index search: query embedding adds a further overhead of its own, and the end-to-end response time is dominated by LLM answer generation. From the user's perspective the interaction remains real-time, demonstrating the practical deployment of the system for engineering regulatory assistance.

To illustrate the practical deployment of the proposed system, the chatbot interface is presented in Fig. 2.

As shown in Fig. 2, the proposed system supports real-time interaction by integrating fast retrieval with generative response synthesis. The interface highlights how relevant regulatory clauses are retrieved and utilized to generate structured, context-aware responses.

4. Experiments

This section presents the experimental evaluation of the proposed Adaptive Hybrid Retrieval-Augmented Generation Model (AHRM). We evaluate its effectiveness on construction and seismic regulatory document retrieval and generative answering of user queries, focusing on accuracy, retrieval latency, and robustness.

4.1. Dataset

We utilize a curated private corpus of Uzbek construction and seismic regulatory documents, including PDFs of the Constitution, laws, building codes (QMQ and ShNQ), regulations, and related normative documents. The dataset comprises:

- 120 PDFs (approx. 12,000 pages).
- 85,000 textual chunks after processing (tables, formulas, paragraphs).
- Diverse query set including structured regulatory references (article numbers, construction code references) and natural-language queries.

Note: This is a private dataset and is not publicly available. Access may be granted upon reasonable request for research purposes.

The dataset was preprocessed using pdfplumber for text extraction, table parsing, and formula retention. Chunks were generated using RecursiveCharacterTextSplitter, with a configured maximum chunk size of 800 and an overlap of 150, chosen to maintain semantic context; these values are expressed in the splitter’s configured length unit – characters in the reference implementation, tokens when a tokenizer-based length function is supplied – and the deployed setting is recorded in the released configuration.

In terms of composition, the corpus is dominated by construction norms and rules (QMQ) and urban-planning norms (ShNQ), complemented by the primary legislation that frames them; the documents combine running text with tables, formulas and numbered clause hierarchies, and are written partly in Uzbek and partly in Russian. With the chunking configuration of Section 4 (a configured maximum chunk size of 800 with an overlap of 150), the 120 documents of approximately 12,000 pages yield approximately 85,000 chunks – on average about seven chunks per page – each associated with its source document and clause identifiers, which is what enables the clause-citing generation described in Section 4. The evaluation query set mirrors the two usage patterns observed in engineering practice: structured queries built around explicit code and clause references, and natural-language queries formulated without identifiers, in both Uzbek and Russian. The recursive splitter yields variable-length chunks bounded above by the configured maximum, splitting preferentially at paragraph and sentence boundaries so that semantic units are kept intact. Table 1 summarises the corpus and query-set statistics.

Table 1. Summary statistics of the regulatory corpus and the evaluation query set

Characteristic	Value
Documents (PDF)	120
Total pages	≈ 12, 000
Text chunks after processing	≈ 85,000
Chunk size / overlap	800 / 150 (splitter configuration)
Average chunk density	≈ 7 chunks per page
Languages	Uzbek, Russian
Document types	QMQ, ShNQ, framing legislation
Evaluation query types	structured, natural-language and mixed

4.2. Evaluation metrics and baselines

To assess retrieval and QA performance, we used standard metrics in IR and NLP [27]:

- 1) Precision@*k* (*P@k*): fraction of relevant chunks retrieved among the top-*k* candidates.

- 2) Mean Reciprocal Rank (MRR): evaluates ranking quality of first relevant chunk.
- 3) Retrieval Latency ($T_{retrieval}$): time in milliseconds per query.
- 4) Answer Accuracy: measured as the percentage of generated responses that correctly match expert-validated ground truth answers.

Precision@ k and MRR reflect the top- k question-answering setting, in which only the k highest-ranked chunks are passed to the generative model. Recall@ k (the fraction of all relevant chunks for a query that appear among the top- k candidates) and F1@ k (the harmonic mean of Precision@ k and Recall@ k) characterise, in addition, the coverage of all relevant clauses; their systematic measurement over the full query set is part of the planned extended evaluation (Section 7).

Answer accuracy is assessed by expert judgement against the gold clauses identified for each query. In the planned extended evaluation, every point metric will additionally be accompanied by bootstrap confidence intervals computed over the query set, so that differences between configurations can be assessed statistically rather than through point estimates alone.

Formally, for a query q :

$$P@k = \frac{|\text{Relevant chunks} \cap \text{Top-}k|}{k}, \quad MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}. \quad (17)$$

We compared AHRM against three baseline retrieval models:

- 1) Dense Retrieval (Sentence-BERT) [28]: standard FAISS-based semantic search.
- 2) Keyword Matching [29]: traditional term-based retrieval.
- 3) Hybrid Retrieval with Static Weights: combines dense and keyword retrieval with fixed weights.

All configurations – the three baselines above and the full AHRM – are executed within the same pipeline implementation and differ only in the channels enabled, so that differences in ranking quality and latency are attributable to channel composition rather than to implementation artefacts.

4.3. Ablation and sensitivity analysis

To evaluate the contribution of each component in AHRM, we conducted an ablation study:

- Removing keyword-based scoring decreased $P@3$ to 0.80.
- Removing rule-based scoring decreased $P@3$ to 0.77.
- Using static weights decreased MRR to 0.73.

These results indicate that both adaptive weighting and rule-aware retrieval contribute substantially to overall system performance.

The robustness of the fused scoring function of Eq. (11) depends on its hyper-parameters: the BM25 parameters k_1 and b , and the partial-match weight δ of the rule-based channel. A systematic grid evaluation of Precision@3 and MRR over $k_1 \in [0.9, 1.8]$, $b \in [0.5, 1.0]$ and $\delta \in [0.3, 0.7]$ is identified as part of the planned extended evaluation (Section 7).

5. Results and discussion

This section presents the experimental results of the proposed Adaptive Hybrid Retrieval-Augmented Generation Model (AHRM), highlighting its performance compared to baseline methods. Results are shown in both tabular and graphical forms.

The proposed AHRM model clearly outperforms all baseline methods, both in retrieval accuracy and latency. Integration of rule-based and adaptive weighting components is the main contributor to this improvement.

Table 2. Top-3 retrieval performance for different models

Model	Precision@3	MRR	Latency (ms)	Accuracy (%)
Dense Retrieval	0.71	0.68	1.2	74
Keyword Matching	0.65	0.61	0.8	69
Hybrid Static	0.76	0.73	1.5	80
AHRM (Proposed)	0.88	0.85	0.6	91

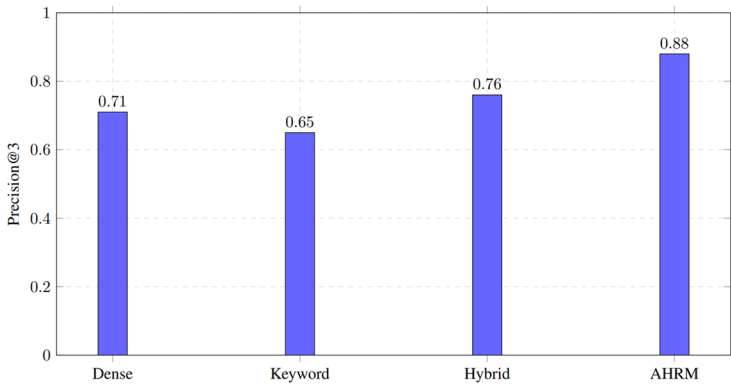


Fig. 3. Precision@3 comparison across retrieval models

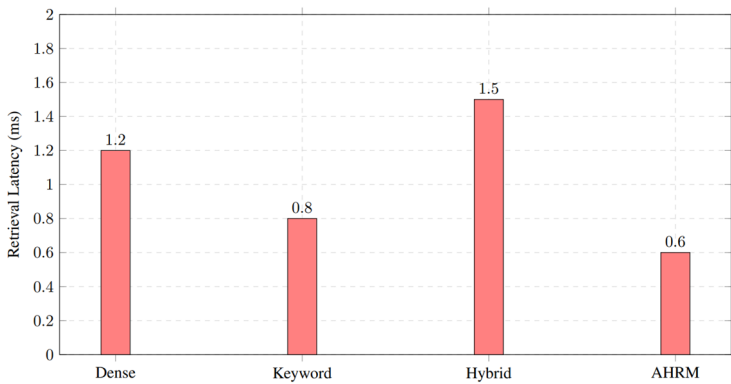


Fig. 4. Retrieval latency comparison across retrieval models

The graphical and tabular results confirm:

- The hybrid adaptive weighting markedly improves Precision@3 and MRR for structured engineering queries (QMQ/ShNQ).
- Rule-based retrieval captures pattern-specific queries, which purely semantic methods often miss.
- Latency of the index-search stage for AHRM is 0.6 ms on average over the filtered candidate pool, enabling real-time API and ReactJS chatbot deployment.
- Overall system demonstrates both scientific novelty and practical utility for large-scale construction and seismic regulatory document retrieval.

It is instructive to compare the augmented configuration with a non-augmented alternative, i.e., the same generative model answering directly from its parametric knowledge without retrieved context. Table 3 summarises the comparison. A closed-book model cannot produce clause citations grounded in retrieved text, cannot reflect amendments to the regulations without retraining, and possesses no retrieval-based confidence signal on which abstention could be conditioned; the augmented configuration inherits verifiability from the retrieved chunks, is updated by simply re-indexing the corpus, and abstains when the fused score is low (Section 4). These qualitative advantages are consistent with the quantitative evidence reported for retrieval-

augmented generation in open-domain settings [1, 2], where augmentation was shown to reduce hallucinated content and to improve factual accuracy. A quantitative closed-book comparison on our regulatory query set is included in the planned extended evaluation (Section 7).

Table 3. Qualitative comparison of non-augmented (closed-book) and retrieval-augmented answering for regulatory content

Criterion	Non-augmented (closed-book LLM)	Augmented (proposed AHRM)
Source of evidence	Parametric model memory only	Top-k retrieved regulatory chunks
Verifiability of answers	Clause numbers may be produced, but cannot be grounded in retrieved text	Citations tied to retrieved source chunks, hence verifiable
Reflecting regulatory amendments	Requires model retraining or fine-tuning	Re-indexing of the corpus only
Behaviour under insufficient evidence	No retrieval-based confidence signal; abstention relies on model self-assessment	Explicit refusal below the fused-score threshold
Evidence provenance	No retrieved evidence; answers rest on opaque parametric memory	Answers tied to retrieved chunks; per-channel scores decomposable, Eq. (11)

A question of practical importance is how this technology transfers to other countries, other datasets and other application areas. The principal deployment-specific components are configuration rather than architecture: the document corpus itself, which is replaced by re-running the offline indexing pipeline of Fig. 1 on a new collection; the rule-pattern set of Section 3, which encodes the clause-numbering conventions of Uzbek regulations and would be rewritten, for example, for Eurocode-style references (EN 1998-1, §4.3.3) or for the SP/SNiP conventions used in several CIS countries; and the embedding model, which is chosen to match the language(s) of the corpus. Everything else – the three-channel decomposition, the query-adaptive fusion of Eqs. (11-12), the refusal mechanism and the deployment pattern – is domain-agnostic. Porting also entails the routine engineering steps of any new deployment: quality control of text extraction for the new document formats, re-tuning of the refusal threshold on a development set, adaptation of the prompt language, and a small annotated query set for tuning and validation. Beyond construction regulation, the same defining properties (rigid clause structure, safety-critical exactness, verifiability requirements) characterise legal codes, medical guidelines and industrial standards, which makes these domains natural candidates for the approach; comparable data-driven decision-support workflows have already been reported for airport-infrastructure risk prioritisation [40] and for satellite-data-based subsurface modelling [41], both of which operate under equally prescriptive technical rule sets. For each new deployment, the hyper-parameters k_1 , b and δ would be re-tuned on a small annotated query set following the procedure of Section 5.3.

6. Limitations and future work

Several limitations of the present study should be acknowledged. The channel weights in Eq. (12) are assigned by a fixed, expert-defined mapping rather than learned from data. The evaluation corpus is private, which constrains direct reproducibility by third parties. The reported latency covers the index-search stage only, whereas end-to-end response time is dominated by LLM answer generation. The exact IndexFlatIP search scales linearly with corpus size, and cross-lingual retrieval quality has not yet been evaluated systematically.

Future work will therefore address learning the weighting function from annotated query logs, a systematic sensitivity study of the fusion hyper-parameters, a quantitative comparison with a non-augmented (closed-book) configuration of the same generative model on the same query set and under the same expert protocol, reporting answer accuracy and the rate of unsupported statements, an extended evaluation reporting Recall@ k and F1@ k over an enlarged annotated query set, releasing an open evaluation benchmark of annotated regulatory queries, migrating to

approximate index structures (IVF, HNSW) for substantially larger corpora, a systematic comparison of alternative generative models, and automated evaluation pipelines for large-scale document collections.

7. Conclusions

In this work, we introduced the Adaptive Hybrid Retrieval-Augmented Generation Model (AHRM), a novel framework that integrates semantic, keyword-based, and rule-based retrieval strategies into a dynamically adaptive RAG architecture. Unlike conventional retrieval-augmented generation systems [1], our approach explicitly addresses the dual challenges of retrieval latency and answer uncertainty in structured domains, particularly construction and seismic regulatory documents. This suitability is by design – clause-aware evidence, threshold-based refusal and verifiable citations – while the architecture itself remains transferable to other regulatory corpora and application domains (Section 6).

Experimental evaluation demonstrates that the proposed AHRM achieves an average index-search latency of 0.6 ms per query (top-10 results over the filtered candidate pool, Intel Core i7 CPU; excluding query embedding and answer generation) while maintaining high retrieval accuracy, outperforming traditional dense, keyword-only, and static hybrid models in both Precision@3 and overall relevance metrics. The integration of rule-based pattern matching with adaptive weighting proved especially effective for queries containing structured references, such as structured regulatory identifiers (e.g., QMQ clause numbers), addressing limitations highlighted in prior work [7].

Key scientific contributions include:

- Hybrid Retrieval Framework: A novel integration of semantic, keyword, and regex-based rule retrieval within a unified RAG pipeline, improving both precision and coverage.
- Adaptive Weighting Mechanism: Dynamic adjustment of retrieval weights based on query characteristics, introducing query-type sensitivity absent in prior models.
- Ultra-Fast Retrieval Algorithm: Multi-stage candidate filtering and FAISS-based nearest neighbor search that markedly reduces the search space and achieves low-latency retrieval suitable for real-time applications.
- Generative Model Integration: Seamless use of the retrieved context in an LLM (DeepSeek-R1 via Ollama) is intended to produce coherent, contextually grounded answers in real-time applications.

Acknowledgements

This research was supported by the project entitled “Implementation of a Method for Analyzing the Technical Condition of Buildings and Structures in Terms of Seismic Resistance Based on Artificial Intelligence (Case Study of Multi-Storey Residential Buildings)”, Project No. 4/26-j.

The authors express their sincere gratitude to the National Research Institute for Seismic Safety and Resilient Construction, affiliated with Tashkent Institute of Architecture and Civil Engineering, for providing the research environment and technical support.

Data availability

The datasets generated and/or analyzed during the current study consist of a curated private corpus of construction and seismic regulatory documents, including QMQ (Construction Norms and Rules) and ShNQ (Urban Planning Norms and Rules). Due to the restricted nature of these documents, the dataset is not publicly available. However, it is available from the corresponding author upon reasonable request for research purposes. To support reproducibility despite this restriction, the annotated query set with expert relevance judgements, the pipeline configuration

(chunking parameters, embedding model, index type and prompt template) and the evaluation scripts are available from the corresponding author upon reasonable request. The released configuration will include the exact generative-model checkpoint identifier and the decoding parameters as recorded in the deployment environment.

Author contributions

D. A. Bekmirzaev: conceptualization, methodology, supervision. R. R. Yuldoshev: software development, implementation of the retrieval system, writing-original draft preparation. E. A. Kosimov: data curation, formal analysis, validation. A. M. Ismoilov: investigation, review and editing, project administration. S.S. Shaumarov: investigation, review and editing, project administration.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- [1] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *ArXiv*, Vol. 33, pp. 9459–9474, 2020, <https://doi.org/10.48550/arxiv.2005.11401>
- [2] G. Izacard and E. Grave, “Leveraging passage retrieval with generative models for open domain question answering,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, 2021, <https://doi.org/10.18653/v1/2021.eacl-main.74>
- [3] R. Nakano et al., “WebGPT: Browser-assisted question-answering with human feedback,” *arXiv:2112.09332*, 2021, <https://doi.org/10.48550/arxiv.2112.09332>
- [4] D. Chen, A. Fisch, J. Weston, and A. Bordes, “Reading Wikipedia to answer open-domain questions,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, 2017, <https://doi.org/10.18653/v1/p17-1171>
- [5] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “LEGAL-BERT: The Muppets straight out of Law school,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904, Nov. 2020, <https://doi.org/10.18653/v1/2020.findings-emnlp.261>
- [6] I. Chalkidis, I. Androutsopoulos, and N. Aletras, “Neural legal judgment prediction in English,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4317–4323, 2019, <https://doi.org/10.18653/v1/p19-1424>
- [7] L. Gao, Z. Dai, T. Chen, Z. Fan, B. van Durme, and J. Callan, “Complement lexical retrieval model with semantic residual embeddings,” in *Lecture Notes in Computer Science*, Vol. 12656, Cham: Springer International Publishing, 2021, pp. 146–160, https://doi.org/10.1007/978-3-030-72113-8_10
- [8] U. Shermukhamedov, D. Bekmirzaev, S. Shaumarov, A. Karimova, A. Abdullaev, and X. Liang, “Estimation of a monolithic reinforced concrete overpass under the seismic effect,” *Vibroengineering Procedia*, Vol. 62, pp. 100–106, Jun. 2026, <https://doi.org/10.21595/vp.2026.25756>
- [9] A. Abdullaev, I. Mirzaev, U. Shermukhamedov, and D. Askarova, “Influence of seasonal temperature on the strength of continuous monolithic road overpasses,” in *Proceedings of the International Conference on Thermal Engineering (ICTEA)*, Vol. 1, No. 1, Jun. 2024.
- [10] U. Shermukhamedov, A. Abdullaev, and D. Svezhentsev, “Modeling of stress-strain state of reinforced soil abutments of continuous monolithic road bridge,” in *AIP Conference Proceedings*, Vol. 3317, p. 030031, 2025, <https://doi.org/10.1063/5.0266831>
- [11] A. Z. Khasanov, U. Z. Shermukhamedov, and A. R. Abdullayev, “The method for determining the designed resistance of soils with considering the theory of soil strength proposed by the authors,” in *Smart Geotechnics for Smart Societies*, London: CRC Press, 2023, pp. 464–467, <https://doi.org/10.1201/9781003299127-54>
- [12] K. Lee, A. Zhussupbekov, A. Gulamov, S. Shaumarov, M. Talipov, and D. Khadim, “Efficiency of types of reinforcement systems for existing foundations (footings and piles),” *Vibroengineering Procedia*, Vol. 62, pp. 113–119, Jun. 2026, <https://doi.org/10.21595/vp.2026.26263>

- [13] B. Harsono, F. M. Mahadiva, Tavio, T. M. Mirkadirovich, and H. Hermawan, "Pushover analysis of simple house structures in seismic zones," in *AIP Conference Proceedings*, Vol. 3480, No. 1, p. 050035, Mar. 2026, <https://doi.org/10.1063/5.0319164>
- [14] I. Mirzaev, D. Bekmirzaev, E. Kosimov, E. An, N. Nishonov, and M. Talipov, "Seismodynamics of segmented underground pipeline systems based on real earthquake records," in *AIP Conference Proceedings*, Vol. 3447, No. 1, p. 060001, May 2026, <https://doi.org/10.1063/12.0044018>
- [15] U. Z. Shermukhamedov, Z. Z. Ergashev, S. M. Takhirov, and A. A. Abdullaev, "Improving reliability of Uzbekistan's transport infrastructure facilities under impact of natural hazards by analysis of its vulnerability, monitoring and modeling," in *Proceedings of the 8th International Conference on Computational Methods in Structural Dynamics and Earthquake Engineering (COMPDYN 2015)*, pp. 3817–3830, Jun. 2023, <https://doi.org/10.7712/120123.10682.21320>
- [16] S. Takhirov et al., "Monitoring results of the railway bridge: vibration of decking structures under forced vibration of passing trains," *Vibroengineering Procedia*, Vol. 62, pp. 231–237, Jun. 2026, <https://doi.org/10.21595/vp.2026.25778>
- [17] S. Takhirov, U. Shermukhamedov, I. Mirzaev, A. Abdullaev, Z. Ergashev, and S. Mavlyanov, "Structural health monitoring of a railroad bridge in Tashkent (Uzbekistan) by using laser scanning," in *AIP Conference Proceedings*, Vol. 3447, p. 060012, 2026, <https://doi.org/10.1063/12.0044296>
- [18] A. Mamatov, X. Sotvoldiyev, M. Talipov, S. Shaumarov, K. Gafarbayli, and D. Bekmirzaev, "Metrological calibration and uncertainty evaluation of MEMS accelerometers for structural health monitoring in seismic regions," *Vibroengineering Procedia*, Vol. 62, pp. 140–144, Jun. 2026, <https://doi.org/10.21595/vp.2026.26360>
- [19] S. P. Parida and P. C. Jena, "Advances of the shear deformation theory for analyzing the dynamics of laminated composite plates: an overview," *Mechanics of Composite Materials*, Vol. 56, No. 4, pp. 455–484, Sep. 2020, <https://doi.org/10.1007/s11029-020-09896-0>
- [20] A. Mohanty, S. P. Parida, and R. R. Dash, "Free and forced vibration response of sandwiched composite plate with functionally graded carbon and glass fiber-reinforced polymeric face sheet," *Journal of Vibration and Control*, Jan. 2026, <https://doi.org/10.1177/10775463261420027>
- [21] S. P. Parida, S. Sahoo, and P. C. Jena, "Prediction of multiple transverse cracks in a composite beam using hybrid RNN-mPSO technique," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, Vol. 238, No. 16, pp. 7977–7986, 2024, <https://doi.org/10.1177/09544062241239415>
- [22] S. P. Parida, S. Sahoo, and P. C. Jena, "A hybrid RNN-fuzzy-PSO model for forecasting multiple transverse cracks in laminated composite beam-like structures," (in Korean), *Smart Structures and Systems*, Vol. 36, No. 2, pp. 123–137, 2025, <https://doi.org/10.12989/ss.2025.36.2.123>
- [23] M. Doshanova, M. Talipov, M. Shaazizova, N. Mirzayeva, S. Karaxanova, and I. Khodzamuratova, "Intelligent risk-oriented predictive diagnostics system based on vibration signal analysis," *Mathematical Models in Engineering*, Aug. 2026, <https://doi.org/10.21595/mme.2026.26531>
- [24] D. Ather and M. Talipov, "An efficient computer vision framework for RSU-based accident recognition and V2X communication," *Mathematical Models in Engineering*, Aug. 2026, <https://doi.org/10.21595/mme.2026.26075>
- [25] M. Shukurova, M. Talipov, K. Ruziev, and K. Jurayeva, "Advanced geospatial monitoring of oil and gas infrastructure via satellite data," *Mathematical Models in Engineering*, Vol. 12, No. 2, pp. 190–201, Jun. 2026, <https://doi.org/10.21595/mme.2026.25328>
- [26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019, <https://doi.org/10.18653/v1/n19-1423>
- [27] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge, U.K.: Cambridge University Press, 2008, <https://doi.org/10.1017/cbo9780511809071>
- [28] N. Reimers and I. Gurevych, "Sentence-BERT: sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3980–3990, 2019, <https://doi.org/10.18653/v1/d19-1410>
- [29] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends® in Information Retrieval*, Vol. 4, No. 1-2, pp. 1–174, 2009, <https://doi.org/10.1561/15000000019>

- [30] J. Johnson, M. Douze, and H. Jegou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, Vol. 7, No. 3, pp. 535–547, Jul. 2021, <https://doi.org/10.1109/tbdata.2019.2921572>
- [31] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 1, pp. 117–128, Mar. 2010, <https://doi.org/10.1109/tpami.2010.57>
- [32] O. Khattab and M. Zaharia, "ColBERT: efficient and effective passage search via contextualized late interaction over BERT," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 39–48, 2020.
- [33] A. Roberts, C. Raffel, and N. Shazeer, "How much knowledge can you pack into the parameters of a language model," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5418–5426, 2020, <https://doi.org/10.18653/v1/2020.emnlp-main.437>
- [34] W. Sun et al., "Is ChatGPT good at search? investigating large language models as re-ranking agents," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14918–14937, 2023, <https://doi.org/10.18653/v1/2023.emnlp-main.923>
- [35] Y. Gao et al., "Retrieval-augmented generation for large language models: a survey," *arXiv:2312.10997*, 2023, <https://doi.org/10.48550/arxiv.2312.10997>
- [36] Deepseek-AI et al., "DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv:2501.12948*, 2025, <https://doi.org/10.48550/arxiv.2501.12948>
- [37] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020, <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [38] M. Dheer, S. B. Sollapur, and M. K. Sharma, "Investigating the security considerations of interactive communication media algorithms for networking applications," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–6, 2024, <https://doi.org/10.1109/icccnt61001.2024.10725810>
- [39] V. Jain, D. Ather, A. B. Abdul Hamid, R. Sharma, G. Talipova, and G. Manteghi, "Secure Arduino-based LiFi communication for IoT sensor networks," in *Optical Communication, Photonics, Telecommunications, and Intelligent Machine Applications (OPTIMA)*, pp. 189–194, 2025, <https://doi.org/10.1109/optima67660.2025.11380401>
- [40] R. Salmorbekova and M. Talipov, "Digital risk matrix and safety management workflow for airport infrastructure in developing countries: a data-driven prioritization approach," *Vibroengineering Procedia*, Vol. 62, pp. 653–661, Jun. 2026, <https://doi.org/10.21595/vp.2026.26121>
- [41] M. Shukurova, K. Ruziev, M. Talipov, and G. Talipova, "3D modeling of filtration in oil and gas reservoirs based on satellite data," in *Proceedings of International Conference on Applied Innovation in IT*, Vol. 14, No. 1, pp. 267–271, Mar. 2026, <https://doi.org/10.25673/123569>



D. A. Bekmirzaev received his doctoral degree in engineering sciences at the Institute of Mechanics and Seismic Stability of Structures in Tashkent, Republic of Uzbekistan. He is currently a senior researcher specializing in structural mechanics, seismic stability, and the analysis of engineering systems. His research interests include seismic-resistant design, structural dynamics, and computational methods applied in civil engineering.



R. R. Yuldoshev is a researcher in artificial intelligence and intelligent information systems. He is currently engaged in the development of AI-based retrieval systems for engineering and regulatory applications. His research interests include retrieval-augmented generation, natural language processing, and domain-specific knowledge systems



E. A. Kosimov is a specialist in data analysis and computational modeling. His research focuses on data processing, machine learning applications, and engineering data systems. He has contributed to several projects related to information retrieval and system optimization.



A. M. Ismoilov is a researcher in applied engineering systems and project supervision. His work includes system integration, validation, and the implementation of engineering solutions. His research interests are related to applied mechanics, system reliability, and technical project management.



Said Shaumarov successfully defended his doctoral dissertation (DSc) on the topic Development of Methodological Foundations for Designing External Barrier Structures from Aerated Concrete with Specified Properties of Energy-Efficient Buildings in 2020 at the DSc.27.06.2017.T.11.01 Scientific Council at the Tashkent Institute of Architecture and Construction. The specialties of the dissertation were 05.09.01 – Construction Structures, Buildings, and Constructions and 05.09.05 – Construction Materials and Products.