Eur. Phys. J. Spec. Top. (2026) 235:8297–8311
https://doi.org/10.1140/epjs/s11734-026-02476-8

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

Self-organizing search on hierarchical normative hypergraphs for decision support in disaster medicine

Oleg E. Karpov[1], Alexander K. Kuc[2], Nikita M. Smirnov[2], Mikhail N. Zamyatin[1], Aleksei V. Osipov[1], Andrey I. Kilnik[1], Semen A. Kurkin[2], Alexander P. Maksachuk[1], and Alexander E. Hramov[1,2,a]

[1] Pirogov National Medical and Surgical Center, 70 Nizhnyaya Pervomayskaya street, Moscow 105203, Russia
[2] Plekhanov Russian University of Economics, 36 Stremyanny Lane, Moscow 115054, Russia

Received 27 May 2026 / Accepted 19 June 2026 / Published online 2 July 2026
© The Author(s), under exclusive licence to EDP Sciences, Springer-Verlag GmbH Germany, part of Springer Nature 2026

Abstract The effective management of disaster medicine requires rapid access to a hierarchical corpus of normative documents—federal regulations, regional adaptations, and local protocols—that often conflict across jurisdictions. Traditional retrieval-augmented generation (RAG) systems rely on pairwise semantic similarity and fail to capture higher-order constraints such as jurisdiction, emergency type, response stage, and actor role. Here we propose a self-organizing RAG architecture built on a hypergraph of 246 507 text chunks derived from 1 239 normative documents of the Russian disaster medicine system. Nodes are chunks embedded with the multilingual model BAAI/bge-m3 (HR@5=0.85), and hyperedges encode shared metadata (region, disaster type, role, response stage). We compare three generation strategies (baseline LLM, standard RAG, and an iterative search agent) on 17 expert-validated queries using LLM-as-a-Judge scoring. Standard RAG achieves the highest average score (2.74 vs. 2.44 for baseline and 2.49 for the agent) but exhibits a fundamental limitation: semantically close chunks from different regions form overlapping attractors in the embedding space, leading to compilative noise and inconsistent answers. To overcome this, we introduce a two-stage self-organizing retrieval pipeline: (i) metadata extraction from the query using the same LLM, and (ii) hypergraph filtering followed by semantic search within the reduced subset. This cascade breaks the symmetry between regional attractors, reduces answer entropy, and ensures normative validity without model retraining. The work demonstrates how hypergraph-based filtering and self-organization principles can enhance AI-driven decision support in hierarchically regulated medical domains, with potential extensions to other fields facing conflicting information sources.

1 Introduction

The mitigation of medical-sanitary consequences of emergencies (disasters) currently represents a complex dynamic system in which the interaction of multiple actors (federal and regional centers, emergency medical services, hospitals) is governed by a hierarchical system of regulatory documents and response algorithms [1]. In the Russian Federation, this system exhibits a pronounced networked hierarchical structure: a central hub (the Ministry of Health of the Russian Federation and the Federal Center for Disaster Medicine as its key operational management body) sets global rules, regional nodes (territorial disaster medicine centers) adapt these rules to local conditions, and local algorithms determine the actions of specific executors [2]. Such a hierarchy generates not only vertical "federal–regional" connections but also horizontal connections between regional documents, as well as referential relationships that form hyperedges of contradictions (for example, when one regional algorithm references a current federal document, while another references an algorithm that has lost its relevance following changes in federal legislation, thereby creating conflicting interpretations) [3].

Traditional systems of decision support systems (DSS) in these conditions demonstrate limited effectiveness as they are unable to account for the multi-level nature of the regulatory framework [4]. Recently, approaches based on large language models (LLMs) and retrieval-augmented generation (RAG) technology have been actively developing [5]. In our previous work, we proposed a hybrid DSS architecture based on a "cognitive digital twin" that integrates dynamic modeling with a domain-adapted LLM (Catastrophe-LLM) [6]. However, classical RAG

[a] e-mail: hramovae@gmail.com (corresponding author)

typically reduces to pairwise semantic search—nearest neighbor search in the vector space of embeddings [7]. In the context of regulatory hierarchy, this is insufficient: fragment relevance is determined not only by its semantic similarity to the query but also by higher-order constraints—jurisdiction (federal/regional level), disaster type, response phase, and executor role (dispatcher, paramedic, manager) [8, 9]. Ignoring these dimensions leads the system to retrieve semantically similar but normatively inapplicable fragments (e.g., an algorithm intended for another region), generating contradictory or unfounded responses [10].

At the same time, the phenomenon of self-organization—the ability of a system to adapt its structure without external control to achieve a stable outcome—is key to complex adaptive systems [11–14]. In the context of RAG, self-organization can manifest as the system's ability to filter regional variations, resolve conflicts between sources, and produce a consensus answer based on implicit priorities, for example, giving priority to federal documents [15]. However, existing RAG architectures lack built-in mechanisms for such self-organization. For instance, in the study by Wong H.S. and Wong T.K. (2026), the Multi-Evidence Clinical Reasoning RAG (MECR-RAG) system designed for triage achieved high accuracy by combining local guidelines (Hong Kong Accident and Emergency Triage Guidelines) with a database of 3 000 real-world emergency department triage cases, but did not account for hierarchical conflicts between sources [11]. In the work by Liu S. et al. (2025), the use of a knowledge graph improved retrieval quality; however, the graph was built on a triage handbook and did not incorporate dynamic filtering by jurisdiction [12]. The closest in spirit is the study by Yazaki M. et al., where RAG improved triage accuracy from 35% to 70% compared to a baseline LLM, but the authors likewise did not address the problem of reconciling heterogeneous regulatory sources [16]. Thus, the task of integrating federal and regional regulatory documents into a unified self-organizing RAG system remains unsolved. It should also be noted that the specific nature of disaster medicine places particular demands on the verifiability of recommendations and the ability to trace their regulatory justification, which makes RAG, with its source citation mechanism, the most appropriate technological solution; however, adaptation of this technology to disaster medicine tasks has not yet been undertaken.

The aim of this work is to propose an architecture for a self-organizing RAG on a hypergraph of disaster medicine regulatory documents, to justify its necessity, and to outline implementation pathways.

In the proposed model, nodes are text fragments (chunks) obtained from 1 239 documents at the federal and regional levels; hyperedges are defined by shared metadata (region, disaster type, executor role). We conduct a comparative analysis of three response generation strategies (baseline LLM, classical RAG, and agentic search) using the LLM-as-a-Judge methodology and show that classical RAG, relying solely on semantic similarity, systematically mixes fragments from different regional jurisdictions, generating interpretation conflicts and increasing response entropy. Based on the obtained results, we substantiate the promise of transitioning to two-stage self-organizing filtering (metadata + semantics), which can potentially provide more stable and normatively justified response quality by breaking the symmetry between regional attractors.

2 Methodology

The overall research pipeline (see Fig. 1) comprises seven sequential stages: (1) construction of the initial corpus of regulatory and methodological documents and response algorithms; (2) extraction of textual content using the Chandra 2 optical character recognition model and the Pandoc converter; (3) removal of administrative headers, approval stamps, and other semantically insignificant information from the text; (4) recursive fragmentation (chunking) of the cleaned text with a target size of 500 tokens and an overlap of 50 tokens; (5) building vector representations of chunks using the multilingual embedding model `BAAI/bge-m3` and constructing a two-level index based on the FAISS library; (6) response generation using three alternative strategies (baseline, augmented generation, agentic search) based on a single `Qwen3-32B` model; (7) quantitative assessment of response quality using the LLM-as-a-Judge methodology with a five-point scale for alignment with reference documents. Below, each stage receives a formal description in terms of complex network theory and self-organization.

2.1 The normative corpus as a multi-level graph

The initial corpus includes 1 239 documents (PDF and DOCX) from the Federal Center for Disaster Medicine (FCDM) and the Territorial Centers for Disaster Medicine (TCDM) of all constituent entities of the Russian Federation. Overall, the dataset comprises standard emergency response algorithms, regulatory legal acts, interagency interaction agreements, and scientific publications.

After the stages of extraction, cleaning, and fragmentation (see Sect. 2.2), the corpus is transformed into a set of text fragments—chunks $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ with $N = 246\,507$. Each chunk c_j is associated with a source document $d_i \in \mathcal{D}$, $|\mathcal{D}| = 1264$, as well as with a set of metadata $\mathcal{M} = \{\text{region, disaster type, executor role, response phase}\}$. The metadata are extracted from the directory structure and file paths, which encode a priori expert knowledge from the FCDM.

We construct a hypergraph $\mathcal{H} = (\mathcal{C}, \mathcal{E})$, where:

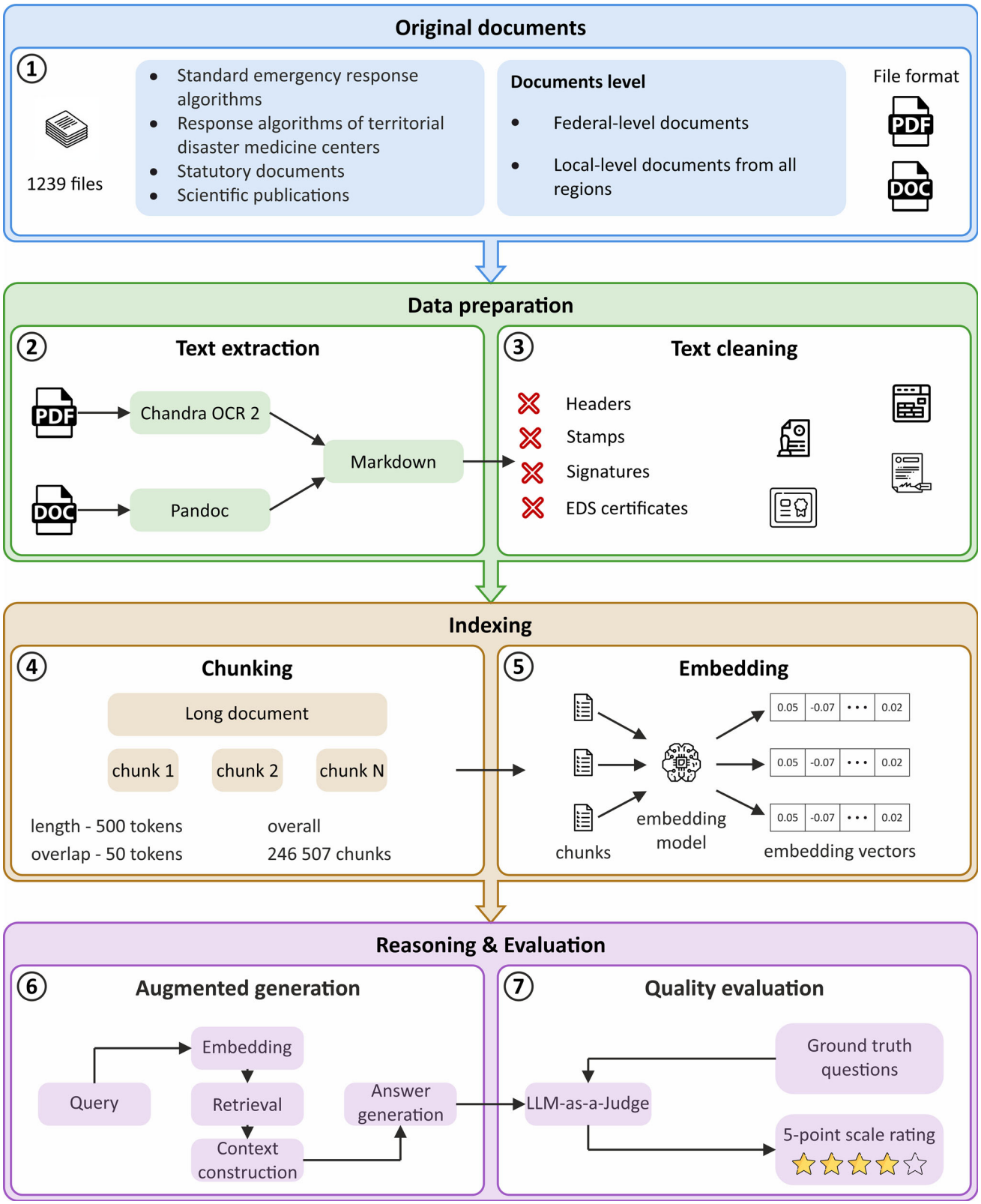


Fig. 1 Pipeline for constructing and evaluating a self-organizing knowledge base on a hypergraph of disaster medicine regulatory documents

- *Nodes* are the chunks c_j;
- *1st-order edges* (pairwise connections) are given by the cosine similarity between the vector representations $\mathbf{x}_j = \mathrm{emb}(c_j)$ (see Sect. 2.3):

$$w_{jk} = \frac{\mathbf{x}_j \cdot \mathbf{x}_k}{\|\mathbf{x}_j\|\|\mathbf{x}_k\|}, \quad w_{jk} \in [0, 1],$$

  where all chunk vectors $\mathbf{x}_j$ are normalized to unit L_2-norm, making their dot product equivalent to cosine similarity. Note that two chunks with similar meanings have vectors that are close in terms of cosine measure (the angle between them is small). For example, a chunk "report to the head of the regional health department" and a chunk "report to the regional health minister" will have a cosine similarity $w > 0.9$; that is, the strength of pairwise connections in the hypergraph is determined by semantic encoding.
- *Higher-order hyperedges* $e \subseteq \mathcal{C}$ unite all chunks that share a common metadata value (e.g., the same region, the same disaster type, or the same executor role).

Thus, the hypergraph encodes not only semantic proximity through pairwise interactions but also, via hyperedges, the normative-functional groupings that are critical for disaster medicine.

We introduce an incidence matrix $\mathbf{H} \in \{0, 1\}^{N \times L}$, where L is the total number of unique metadata combinations (considering all dimensions). An element $H_{j,\ell} = 1$ if chunk c_j belongs to hyperedge e_ℓ. This matrix subsequently allows filtering of the search space according to metadata specified by the user or automatically extracted from the query.

2.2 Text extraction and fragmentation

Text extraction from PDF documents was performed using the Chandra 2 model (5 billion parameters, Qwen3.5 architecture)—a multimodal "vision-language" model designed to preserve the two-dimensional topology of tabular data. DOCX documents were converted using the `pandoc` tool. Administrative headers ("APPROVED BY", "AGREED UPON", signatures) were removed from the cleaned text using regular expressions.

The stage of splitting text into fragments (chunks) directly determines the quality of subsequent retrieval, since each chunk serves as the minimal extraction unit when processing a query in the RAG pipeline.

The problem of choosing an optimal chunking strategy lies in a fundamental trade-off between two competing requirements. On the one hand, chunks must be sufficiently small to ensure high retrieval accuracy. On the other hand, chunks must be sufficiently large to preserve contextual coherence. Breaking a logically unified fragment leads to a loss of semantic integrity and, consequently, to the generation of incomplete or incorrect responses [17].

In the present study, this trade-off is resolved by applying a recursive text splitter that implements a hierarchical approach to document segmentation [18]. The algorithm operates on an ordered set of separators ranked in descending order of their semantic significance. Let a text T be given along with an ordered set of separators $\Sigma = (\sigma_1, \sigma_2, \ldots, \sigma_m)$ in descending order of semantic significance: σ_1—double line break (paragraph boundary), σ_2—single line break, σ_3—period with space (sentence boundary), σ_4—semicolon, σ_5—comma. For each level i, we define a splitting operator $\mathrm{split}_{\sigma_i}(T)$ that divides T into substrings using the separator σ_i. If the resulting fragments do not exceed the target length $L_{\max} = 500$ tokens, the procedure terminates. Otherwise, for each overly long fragment, the next-level separator σ_{i+1} is applied recursively. Overlap between adjacent chunks is fixed at 50 tokens and is ensured by appending $\delta = 50$ tokens from the end of the previous chunk to the beginning of the next one, which preserves context at segmentation boundaries. The tokenized length of a text fragment is estimated based on an empirically derived coefficient for Russian language texts, according to which the number of tokens is approximated as 2/3 of the total character length of the string.

This configuration ensures that the algorithm primarily strives to preserve the integrity of paragraphs and logical blocks, and only performs splitting at the sentence and syntagm level when necessary. As a last resort, if the text cannot be split using the specified separators, forced truncation is applied at a fixed number of characters corresponding to the maximum token limit.

The chunk corpus was filtered using a text quality metric defined as the proportion of valid characters (alphabetic, numeric, and punctuation marks) in the total string length. Chunks for which this metric fell below a threshold of 0.7 were excluded from further consideration as optical recognition artifacts containing a significant amount of irrelevant or corrupted character information. The final size of the annotated chunk corpus was 246 507 text fragments, with an average chunk size of 282 tokens.

2.3 Vector representations and embedding model

An important stage in constructing a vector knowledge base is the selection of an optimal embedding model—a function that maps discrete text fragments into a continuous vector space of fixed dimensionality. The quality of

this mapping directly determines the effectiveness of subsequent semantic search, since in the RAG pipeline the relevance of the retrieved context is evaluated through a measure of similarity between the vector representation of the query and the vector representation of the chunks [19].

The problem of selecting an embedding model in the present study is complicated by two factors. First, the domain of disaster medicine is characterized by highly specialized Russian language terminology, encompassing both medical concepts and specific regulatory-legal constructs. Publicly available multilingual embedding models trained on general-purpose corpora may demonstrate reduced effectiveness when processing domain-specific vocabulary. Second, there are no standardized benchmarks for quantitatively assessing embedding quality for disaster medicine tasks, making it impossible to directly compare models based on existing reference datasets.

To overcome these limitations, the present study develops a comparative evaluation of embedding models based on generating a synthetic test set of questions using a local generative language model. This approach makes it possible to create a closed validation loop relevant to the task of Russian language text recognition and to perform quantitative comparison of models using unified metrics. It is important to emphasize that this stage does not evaluate the quality of final responses for the user nor does it address medical issues—the procedure is aimed solely at determining the ability of the embedding model to correctly match a query with the text fragment from which it was generated.

The test dataset was constructed in several sequential steps. A truncated sample of 1 000 random fragments, representing the diversity of document types and thematic categories, was extracted from the overall corpus of chunks. From this sample, 20 random chunks were selected for test query generation. Each selected chunk was fed into the local generative model `Qwen3-32B` with a prompt instructing it to generate a question in Russian that is semantically close to the chunk's content, reflects the key concepts from the text, and is sufficiently specific to uniquely identify the given fragment. Generation was performed with parameters ensuring deterministic outputs (temperature set to zero), which guaranteed reproducibility of the results. For each generated question, the source chunk identifier, its textual content, and the path to the original document were recorded, thereby forming a reference test set in which each query has a single known correct match.

The evaluation of embedding model quality was performed according to the following scheme. For each model under test, embeddings were computed from the truncated sample of chunks, after which a vector database indexing this set was constructed. For each generated question, its embedding was computed using the same model, followed by a search for the five most semantically similar chunks in the constructed database. The obtained results were compared against the reference source chunk identifier, making it possible to determine whether the correct chunk was found and at what position in the ranked list it appeared.

For quantitative evaluation, two metrics were used:

- *Hit Rate@K (HR@K)* is the proportion of queries for which the correct chunk (question source) appeared among the top K search results.
- *Mean Reciprocal Rank@K (MRR@K)* is the arithmetic mean of the reciprocal ranks of the first correct chunk:

$$\mathrm{MRR@K} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{\mathrm{rank}_q},$$

where $Q = 20$ is the number of queries, and rank_q is the position of the first correct chunk (if the chunk is not found in the top-K, $1/\mathrm{rank}_q = 0$). Unlike Hit Rate, this metric accounts not only for whether the correct chunk appears in the results but also for its position in the ranked list, penalizing models that place the relevant fragment at lower positions.

In hypergraph terms, HR@K characterizes the model's ability to preserve local structure—associating a query with a semantically close node—while MRR@K reflects the ranking ability required for ordering neighbors. In the present study, the parameter K was set to 5.

The `BAAI/bge-m3` model achieved HR@5 = 0.85 and MRR@5 = 0.70, significantly outperforming the nine other models tested (see Table 1). This model, developed specifically for multilingual tasks and supporting over one hundred languages, demonstrated not only a high ability to find relevant chunks but also excellent result ranking. This means that in 85% of cases, the correct chunk is among the top-5, and on average it appears between positions 1 and 2 (since $1/1.43 \approx 0.7$) in the ranked list. This is critically important for RAG systems, as the context provided as input to the LLM is typically limited to the first few search results. The significant margin in MRR over the models in second and third place (0.56 and 0.52, respectively) confirms the superiority of `BAAI/bge-m3` specifically in terms of ranking ability, rather than merely in its capacity to detect relevant fragments within a broad search radius. Consequently, the `BAAI/bge-m3` model was selected for obtaining vector representations of the chunk corpus.

Table 1 Results of evaluation of the effectiveness of embedding models

No.	Model	HR@5	MRR@5
1	`BAAI/bge-m3`	0.85	0.70
2	`Snowflake/snowflake-arctic-embed-l-v2.0`	0.75	0.56
3	`intfloat/multilingual-e5-large-instruct`	0.75	0.52
4	`sentence-transformers/LaBSE`	0.70	0.56
5	`BAAI/bge-large-en-v1.5`	0.55	0.36
6	`cointegrated/LaBSE-en-ru`	0.50	0.39
7	`thenlper/gte-large`	0.50	0.35
8	`sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`	0.45	0.40
9	`sentence-transformers/paraphrase-multilingual-mpnet-base-v2`	0.40	0.35
10	`sentence-transformers/distiluse-base-multilingual-cased-v2`	0.40	0.34

2.4 Two-level index architecture

To ensure efficient semantic search over the constructed chunk corpus, a two-level vector index architecture based on the FAISS library was implemented. The choice of this library is motivated by its high performance when working with dense vector representations and its support for efficient approximate search algorithms [20].

Micro-level (chunks): An index of type `IndexFlatIP` implements exact nearest neighbor search using the inner product metric. As noted above, all chunk vectors $\mathbf{x}_j$ are normalized to unit L_2-norm, making the inner product equivalent to cosine similarity and allowing search results to be interpreted directly as a measure of semantic proximity.

The lower level of the index aggregates 246 507 vectors corresponding to individual chunks obtained from the recursive segmentation procedure.

Macro-level (documents): For each document $d_i \in \mathcal{D}$, an initial fragment of 12 000 characters is extracted, and its embedding $\mathbf{y}_i$ is computed using the same `BAAI/bge-m3` model. A separate FAISS index stores $M = 1264$ document vectors.

The procedure for extracting relevant information in the multi-level regulatory graph was organized according to a hierarchical scheme.

1. For a query ℓ, its embedding $\mathbf{x}_\ell$ is computed.
2. A search for the $S = 5$ nearest chunks is performed in the micro-index.
3. The retrieved chunks are grouped at the macro level by source document identifier, and all relevant fragments are aggregated for each document. If necessary, the full document vector can be extracted from the macro-index to assess global relevance.

This approach offers a dual advantage: high search granularity at the chunk level ensures precision in localizing relevant information, while subsequent aggregation by document provides the end user or generative model with the complete regulatory context necessary for correct interpretation of the retrieved provisions and verification of their applicability.

2.5 Response generation strategies as modes of self-organization

Three strategies were compared, all implemented on the same generative model—`Qwen3-32B` (32 billion parameters, local deployment). Each strategy is interpreted as a self-organization mode of the system within the hypergraph $\mathcal{H}$.

Baseline Strategy (LLM only) The model receives only the user query ℓ, without any context from the regulatory document corpus. This configuration relies solely on the parametric knowledge embedded in the LLM's weights during pre-training [21]. Since the disaster medicine domain is weakly represented in general-purpose training corpora, in the absence of external relevant context, the generation dynamics in semantic space exhibit a tendency to converge to stable but often irrelevant semantic attractors—a characteristic state for systems with iterative transformations [22, 23]. This feature, in the absence of navigation through an external knowledge base, explains the observed high entropy of responses and the tendency toward "hallucinations"—the generation of factually unreliable information [24].

This configuration serves as a reference point, enabling quantitative assessment of the contribution of information retrieval mechanisms to the final response quality.

Augmented Generation based on Vector Search (RAG) Classical RAG architecture: the query ℓ is converted into an embedding $\mathbf{x}_\ell$, and a search is performed for the $S = 5$ nearest chunks in the micro-index (by cosine similarity). The retrieved chunks are concatenated and fed into the LLM together with ℓ as contextual augmentation. This mode corresponds to fixation on the five nearest nodes in Euclidean space. This approach ensures response relevance by explicitly providing the model with domain-specific information that is absent or underrepresented in its training data.

It is more stable than the baseline but sensitive to regional noise: if the five nearest chunks include fragments from different constituent entities of the Russian Federation, the model may generate a contradictory response.

Search Agent The third strategy represents an agentic approach, in which the language model functions as an autonomous agent endowed with the ability to interact instrumentally with the document corpus [25]. The agent is provided with access to three specialized tools: (1) full-text search over the corpus using the `ripgrep` utility, which enables high-performance regular expression search with context line output; (2) reading an arbitrary fragment of a file specified by file path and line number range; (3) file search by name pattern, allowing localization of documents matching a given mask. The key distinction of this strategy from RAG-based vector search is the iterative nature of the process: the agent independently formulates search queries based on analysis of the current task, interprets the obtained results, determines the need for additional information, and can initiate subsequent search and reading cycles. The interaction protocol limits the maximum number of iterations to fifteen, after which the agent must produce a final answer. This architecture models the behavior of an expert sequentially studying the regulatory framework to find relevant provisions.

From the perspective of dynamics on the hypergraph, this is a walk with feedback: each step of the agent selects a new edge or node based on the previous result. This regime is bistable—with a successful initial hypothesis (the first query hits the correct document), the agent achieves high accuracy, but with an unsuccessful first step, it may drift into a region of irrelevant nodes, resulting in low response quality. The high sensitivity to initial conditions explains the polarized distribution of scores (see Sect. 3).

2.6 Quality assessment procedure: LLM-as-a-judge as an external benchmark

For quantitative comparison of the strategies, the "LLM-as-a-Judge" methodology was employed, using the same `Qwen3-32B` model but in classification mode (without generating new content). The choice of this approach is motivated by the absence of standardized reference datasets for regulatory-legal retrieval tasks in the given domain, as well as the high labor intensity of manual expert evaluation of generated text responses. The LLM-as-a-Judge methodology makes it possible to automate the evaluation procedure while maintaining a high degree of correlation with human expert judgments, as confirmed by a number of independent studies in the field of generative model evaluation [26].

In our study, we also supplemented the LLM-as-a-Judge methodology with an evaluation of the generated responses by FCDM experts. All responses produced by the language model were examined by eight experts—operational duty officers of the disaster medicine service—who assessed response quality in terms of usefulness and importance for the center's current operations.

A key feature of the implemented evaluation scheme is the independent assessment of each generated response against a set of reference documents. This approach provides a granular evaluation of response relevance to each potential information source and makes it possible to identify both the completeness of regulatory framework coverage and the strategy's ability to focus on the most authoritative documents.

Reference Document Set The evaluation set included 17 questions on disaster medicine topics (see Table 3), covering key aspects of professional activity: criteria for classifying events as emergencies, standard response algorithms for various disaster categories, principles of organizing medical care for victims, as well as issues of regulatory-legal regulation of disaster medicine service operations.

For each of the 17 test questions, a set of reference documents $\mathcal{R}_\ell \subseteq \mathcal{D}$ was defined that are known to contain the normatively correct answer. Additionally, documents retrieved by the RAG strategy were included in $\mathcal{R}_\ell$ (to avoid bias in favor of any single strategy). The average size of the reference set per question for the evaluation set was $|\mathcal{R}_\ell| = 7.3$.

Rating scale: For each pair "generated response a—reference document $r \in \mathcal{R}_\ell$", the judge model assigns a score from 1 to 5:

- **5 points**—complete and exact correspondence of the response to the content of r.
- **4 points**—most of the information is reflected, but with minor omissions.
- **3 points**—partial reflection with significant gaps.
- **2 points**—weak relevance (the response only indirectly addresses the topic).
- **1 point**—no correspondence or factual errors.

Since the scale is ordinal, the overall quality of a strategy for question ℓ is defined as the arithmetic mean of the scores over all $r \in \mathcal{R}_\ell$:

$$\text{Score}_\ell = \frac{1}{|\mathcal{R}_\ell|} \sum_{r \in \mathcal{R}_\ell} \text{LLM-judge}(a, r).$$

Such an averaged score can be viewed as a discretization of a continuous measure of semantic agreement (e.g., cosine similarity between the response embedding and the reference fragment), but with consideration of normative correctness, which provides higher validity.

The user message included the original question, the text of the generated response, and the content of the reference document. In cases where the reference document volume exceeded the context window of the judge model, truncation was applied to the first 8 000 characters—a value empirically determined to be sufficient for preserving the semantic integrity of the regulatory document while respecting technical constraints. The judge model returned the result in a structured JSON format containing a numerical score on a five-point scale and a textual justification for the decision, ensuring both quantitative measurability and qualitative interpretability of the results.

In total, $3 \times 124 = 372$ independent judgments were obtained for the three strategies (17 questions $\times$ 7.3 documents on average). The statistical significance of differences was tested using nonparametric methods (results are presented in Sect. 3).

3 Results

A comparative analysis of three response generation strategies (baseline LLM, RAG, and agentic search) was conducted on an evaluation set of 17 questions covering key aspects of disaster medicine. For each question, a set of reference documents $\mathcal{R}_q$ was defined (on average $|\mathcal{R}_q| = 7.3$). The quality of each generated response was assessed using the LLM-as-a-Judge method and by FCDM experts on a five-point scale. The total sample size comprised 124 independent evaluations per strategy (372 judgments in total).

Comparison of Strategies: Mean Scores and Distribution

Table 2 presents the mean scores and score distributions for the three strategies. The RAG strategy achieved the highest mean score (2.74 for LLM-as-a-Judge and 3.46 for expert evaluation), outperforming the baseline LLM (2.44 for LLM-as-a-Judge and 3.05 for expert evaluation) and the agentic search (2.49 for LLM-as-a-Judge and 2.68 for expert evaluation). The proportion of high-scoring responses (4 and 5 points) in RAG was 14.5%, compared to only 7.3% for the baseline. This observation naturally reflects the limited parametric knowledge of the model in the highly specialized domain of disaster medicine. At the same time, the proportion of low-scoring responses (1 – 2 points) in RAG decreased to 25.8% from 38.7% for the baseline model, while the proportion of high scores increased to 14.5%. A fundamentally important result is the appearance, within this strategy's outputs, of responses receiving the maximum score, demonstrating the ability of the RAG approach to ensure complete correspondence between the generated response and the content of the relevant regulatory document. This finding confirms the effectiveness of vector search as a mechanism for providing domain-specific context to the model.

Table 2 Score distribution by strategy (124 evaluations per strategy)

Strategy	Mean score (FCDM experts)	Mean score (LLM-as-a-Judge)	1 point	2 points (%)	3 points (%)	4 points (%)	5 points (%)
Baseline (LLM only)	3.05	2.44	25.0	13.7	54.0	7.3	0.0
Augmented generation (RAG)	**3.46**	**2.74**	16.1	9.7	59.7	12.9	1.6
Agentic search	2.68	2.49	25.8	20.2	39.5	8.1	6.5

The bold values in each column indicate the strategy that achieved the highest percentage of scores for that particular score level (1 through 5)

Table 3 Results by question (mean score)

No.	Question	Baseline	RAG	Agent	Best
1	Criteria for classifying a traffic accident as an emergency	2.3	**3.5**	1.2	RAG
2	Number of emergency medical teams for a mass traffic accident	**3.3**	2.7	2.9	Baseline
3	Dispatcher actions during an explosion	1.8	**2.6**	2.4	RAG
4	Algorithm for natural wildfires	2.3	3.1	**3.6**	Agent
5	Actions during a radiation accident	2.8	**3.1**	2.5	RAG
6	Algorithm for a terrorist attack	2.2	**2.7**	1.8	RAG
7	Algorithm for a sanitary-epidemiological emergency	2.3	**2.6**	2.3	RAG
8	Organization of the All-Russian Disaster Medicine Service during emergencies	3.6	3.6	**3.8**	Agent
9	Completion of accounting forms	2.4	3.0	**3.5**	Agent
10	Information interaction between EMERCOM and the Ministry of Health	3.5	**3.5**	3.5	RAG
11	Medical evacuation during emergencies	2.8	**2.8**	2.1	RAG
12	Actions of an emergency medical paramedic	3.0	**3.6**	3.1	RAG
13	Patient routing during traffic accidents	3.0	2.8	**3.0**	Agent
14	Duties of the TCDM operational duty officer	**2.0**	1.3	1.3	Baseline
15	Reserve of material resources	1.7	**2.8**	1.3	RAG
16	Medical triage during mass casualties	**2.7**	1.7	1.0	Baseline
17	Reporting of infectious morbidity information	1.2	2.2	**3.8**	Agent

The bold values in each row indicate the strategy that achieved the highest average score for that specific question

The agentic search exhibits a polarized distribution: the maximum proportion of highest-quality responses (5 points—6.5%) is combined with the highest proportion of low scores (46.0%). This indicates a bistable operating regime of the agent, sensitive to the initial conditions of the search.

Detailed Analysis of Test Questions

Table 3 presents the mean scores for each of the 17 questions.

The augmented generation strategy achieved the best result for 10 out of 17 questions (58.8% of cases), confirming its advantage as the most robust approach. The superiority of this strategy is particularly pronounced for questions requiring precise reproduction of regulatory criteria and standardized procedures.

The agentic search demonstrated superiority for 4 out of 17 questions (23.5% of cases), with the nature of these questions indicating specific advantages of the agentic approach. The agentic approach possesses a competitive advantage when processing queries that require navigation through hierarchical or corpus-distributed information structures, as well as when integration of information from several thematically related documents is necessary.

The baseline strategy (without augmentation) proved best for only 3 out of 17 questions (17.6% of cases). This phenomenon can be explained by the fact that these questions address basic principles of medical care organization, which are sufficiently well represented in the general training corpus of the `Qwen3-32B` model. In these cases, retrieving fragments from the specialized corpus apparently introduced redundant or contextually biased information, leading to a reduction in the relevance of the final response.

It is important to note that the evaluation of the generated responses by FCDM experts generally coincided with the scores assigned by LLM-as-a-Judge: the mean score determined by experts was 3.05 for the baseline LLM, 3.46 for RAG, and 2.68 for the agent. The highest mean score was assigned to responses generated using RAG; however, when using this strategy, minor omissions were noted for some questions and, in several cases, significant gaps.

The main reason for the experts lowering the RAG evaluation is that, when answering questions that were general in nature and applicable to all constituent entities of the Russian Federation, this strategy generated responses in 40% of cases not based on federal standard algorithms, but on the regulatory documentation of an arbitrarily selected constituent entity of the Russian Federation. In their conclusions, the experts suggested that after data ranking and the introduction of geo-markers, the percentage of correct responses should increase.

It is worth noting that the chosen comparative evaluation method may have underestimated the quality of the agentic search's responses, since in most cases the agent suggested clarifying the question or continuing the discussion of the topic. As a consequence, in 3 cases, according to the experts, the agentic search's response did

not correspond to the question or was weakly related to it (1 – 2 points), and in 5 tasks, the operator chose not to extend the dialog when working with the agent.

Response Entropy as a Measure of Uncertainty

To quantitatively assess the variability of response quality for each strategy, we compute the Shannon entropy of the score distribution on the five-point scale:

$$H = -\sum_{i=1}^{5} p_i \log_2 p_i,$$

where p_i is the proportion of responses receiving i points. Entropy characterizes the degree of uncertainty (or "dispersion") of the results: low values correspond to a concentration of scores within a narrow range, while high values correspond to a uniform distribution across different quality levels.

Based on the distributions presented in Table 2, the following values were obtained: baseline LLM: $H = 1.65$ bits; augmented generation (RAG): $H = 1.67$ bits; agentic search: $H = 2.05$ bits. The baseline LLM and RAG demonstrate nearly identical, moderate entropy. This is explained by the similarity of their distributions: in both cases, medium scores (3 points) dominate, with a relatively small proportion of extreme values (1, 2, 4, 5). Thus, from the perspective of quality predictability, RAG offers no advantage over the baseline model—both strategies are equally "stable" at the level of "mediocre" responses.

In contrast, the agentic search is characterized by a markedly higher entropy (2.05 bits vs. $1.65 \div 1.67$). This is a direct consequence of the bistability of the agentic regime: the agent can with equal probability either fail (high proportion of scores 1 and 2—46.0%) or achieve high quality (the highest proportion of top scores among all strategies—6.5%). High entropy reflects sensitivity to initial conditions and chaotic transitions between different semantic attractors—a property not observed in either the baseline LLM or RAG.

Thus, entropy serves as an indicator of the dynamic regime: low entropy (baseline LLM and RAG) corresponds to a stable, though not necessarily optimal, attractor; high entropy (agent) corresponds to chaotic wandering between several semantic attractors, which is consistent with the polarized score distribution in Table 3.

4 Discussion

Search Quality on the Normative Hypergraph$\mathcal{H}$ under Different Response Generation Strategies

The augmented generation strategy based on vector search provides a stable improvement in response quality compared to the baseline configuration. The mean score on the five-point scale increased from 2.44 to 2.74, while the proportion of responses receiving the lowest score (1 point) decreased from 25.0% to 16.1%. At the same time, there was an almost twofold increase in the proportion of high scores (4 and 5 points)—from 7.3% to 14.5%. These indicators demonstrate the fundamental ability of the RAG approach to compensate for the deficit of domain-specific knowledge in the generative model by providing relevant context from a specialized corpus.

The greatest positive effect from applying augmented generation was observed for questions requiring precise reproduction of quantitative indicators, regulatory criteria, and procedural details contained in governing documents. Such questions include: criteria for classifying a traffic accident as an emergency (increase from 2.3 to 3.5 points), dispatcher procedures upon receiving information about an explosion (increase from 1.8 to 2.6 points), instructions for completing disaster medicine service accounting forms (increase from 2.4 to 3.0 points), as well as the reserve of material resources for emergency response (increase from 1.7 to 2.8 points). In all of the above cases, the relevant information is strictly regulatory in nature and is highly likely to be absent from the training data of general-purpose generative models, which explains the pronounced superiority of the strategy with access to an external knowledge base.

Regional Semantic Attractors

At the same time, for 4 out of 17 questions (23.5% of cases), the application of augmented generation led to a deterioration in response quality compared to the baseline strategy. A detailed analysis of these cases made it possible to identify a characteristic mechanism of error occurrence—insufficient contextualization of the retrieved fragments.

Visualization of vector representations of regulatory document chunks using the t-SNE method (Fig. 2) reveals important features of the geometry of the semantic space formed by the embedding model. The figure shows projections of two subsets of the corpus: regional documents (panel A) and federal documents (panel B).

Analysis of the point distribution shows that, despite the difference in sources, the embeddings of federal and regional documents do not form isolated regions but occupy overlapping subspaces. Regional documents are characterized by the formation of a dense, quasi-continuous cloud without a clearly defined cluster structure, which reflects the high variability of local formulations while maintaining overall semantic proximity. At the same time,

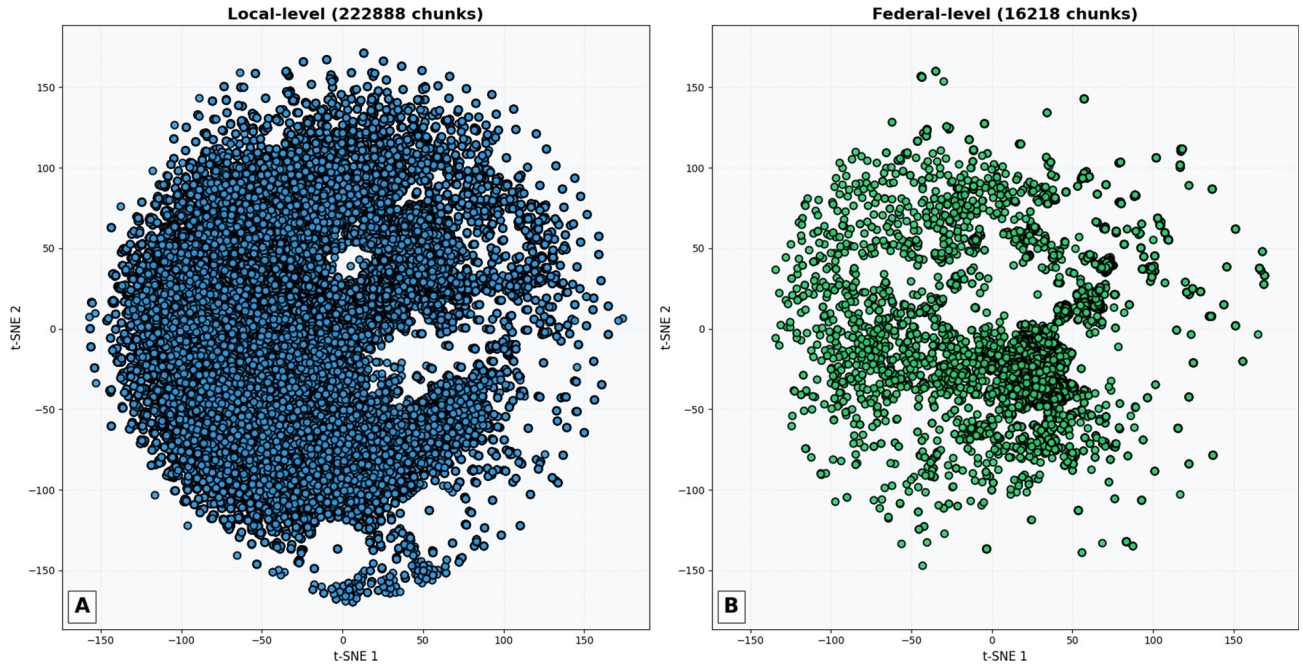


Fig. 2 Visualization of vector representations (embeddings) of text chunks from disaster medicine regulatory documents using the t-SNE method

federal documents exhibit more pronounced local clustering. However, these clusters are not separated from the regional distribution but rather overlap with it.

This effect indicates that the embedding vector space is organized not as discrete semantic regions corresponding to levels of the regulatory hierarchy, but as a system of overlapping attractors. Federal and regional documents form semantically similar but normatively distinct regions that prove to be topologically inseparable in the embedding space.

From a practical standpoint, this means that the classical semantic search mechanism in RAG systems, based on a proximity metric of vector representations, is incapable of reliably distinguishing between sources at different levels of the hierarchy. As a result, the system may retrieve semantically relevant but normatively irrelevant fragments (e.g., regional algorithms for a query that implies the federal level of regulation), leading to a reduction in the validity and reproducibility of responses. For example, for a query formulated in general terms, the system may retrieve a regional algorithm (e.g., for Krasnodar Krai) that is semantically close but legally inapplicable in another constituent entity of the Russian Federation (e.g., in Krasnoyarsk Krai). FCDM experts noted this problem when testing the RAG strategy in three cases, as well as the need for its modification, which will be discussed further in the section.

This phenomenon explains the paradoxical result when the baseline LLM (relying on parametric memory) outperforms RAG for questions with strong regional variability (questions 2, 14, 16 in Table 3). Parametric memory contains averaged federal norms, whereas RAG is "drawn into" a local attractor, retrieving contradictory fragments.

Hyperedges of Contradictions: The Mechanism of Regional Incompatibility

When the system retrieves $S = 5$ nearest chunks, and these chunks belong to five different constituent entities of the Russian Federation, an inconsistent hyperedge is formed—a set of nodes that are semantically close (cosine similarity > 0.8) but normatively incompatible. Receiving such a context, the LLM cannot select a dominant source and generates compilative noise—a response containing averaged or mutually exclusive instructions.

This mechanism is exacerbated by the high degree of information duplication in the corpus. Approximately 90% of the documents are regional TCDM algorithms, which often repeat each other with minor variations in phrasing while preserving an identical semantic core. Consequently, the retrieval system returns five semantically similar fragments originating from documents of different constituent entities of the Russian Federation. Upon concatenation, these fragments form an internally contradictory context containing mismatched numerical values or non-identical lists of measures. Faced with such contradictions, the model is unable to determine a priority source and generates a compilative response that does not fully correspond to any of the regional regulations.

Thus, classical RAG without metadata filtering systematically suffers from the effect of "mixing of regional semantic attractors", which leads to quality degradation in tasks where regulatory applicability depends on jurisdiction.

Limitations of Classical RAG Architecture for Search Tasks on a Normative Hypergraph

The identified limitations are fundamental in nature and point to a fundamental problem inherent in classical RAG architecture when applied to regulatory-legal retrieval tasks: semantic similarity of vector representations does not guarantee relevance according to criteria of regulatory applicability. Two text fragments demonstrating high cosine similarity may belong to different regional jurisdictions, regulate different types of emergencies, or describe action algorithms for different response phases. Without explicit specification of these characteristics, the retrieval system returns any of these fragments with equal probability, which in some cases leads to the generation of normatively unjustified or contradictory recommendations.

Proposed Two-Stage Architecture: A Self-Organizing Filter on a Hypergraph

To overcome the described limitations, we propose replacing the single-stage semantic search with a cascading self-organizing procedure that includes metadata filtering, as shown schematically in Fig. 3.

Step 1: Query parsing and metadata extraction. The user query ℓ is fed into the same LLM `Qwen3-32B` with a prompt instructing it to extract relevant metadata: $\mathcal{M}_\ell$ = {region, disaster type, executor role, response phase}. For example, for the query "Actions of a paramedic during a traffic accident in Moscow Oblast", the extracted metadata are: region = Moscow Oblast, role = emergency paramedic, disaster type = traffic accident.

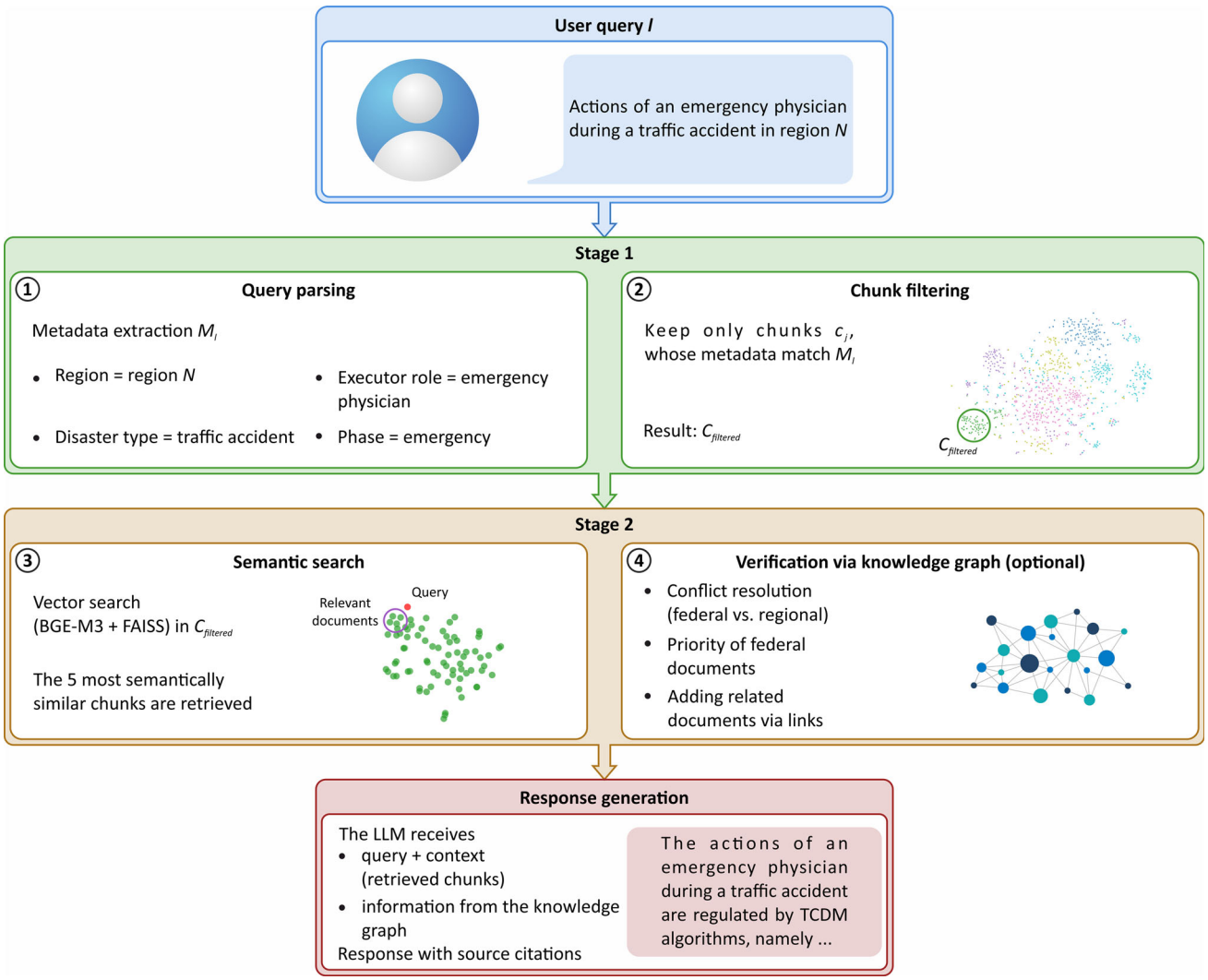


Fig. 3 Schematic diagram of the two-stage self-organizing RAG architecture on a hypergraph of regulatory documents

Step 2: Hypergraph filtering. We introduce a projection operator $P_{\mathcal{M}}$ that maps the query into the metadata subspace:

$$P_{\mathcal{M}}(\ell) = \{c_j \in \mathcal{C} \mid \forall m \in \mathcal{M}_\ell,\ \ H_{j,\,\ell(m)} = 1\},$$

where $\ell(m)$ is the index of the hyperedge corresponding to metadata m. In other words, all chunks whose metadata do not match the extracted metadata are excluded from consideration. This breaks the symmetry between regional attractors: if the query explicitly specifies a region, the system remains only within its basin; if no region is specified, the federal level (global attractor) is selected by default.

Step 3: Semantic search on the reduced subset. In the filtered subset $\mathcal{C}_{\text{filtered}} = P_{\mathcal{M}}(\ell)$, standard vector search is performed using `BAAI/bge-m3` embeddings and the FAISS index. Since the size of $\mathcal{C}_{\text{filtered}}$ is substantially smaller than the total $N = 246507$ (on the order of thousands of chunks for a specific region), the search remains efficient, and the risk of retrieving "foreign" regional norms is eliminated.

Step 4 (optional): Verification via knowledge graph. Further development of the presented methodology is envisioned through the integration of the described two-stage retrieval procedure with a formalized knowledge graph of the subject domain. While enriching chunks with metadata solves the problem of contextual filtering at the level of document attributes, a knowledge graph can explicate semantic relationships between concepts, regulatory acts, and operational entities, thereby opening up possibilities for logical inference and multi-step reasoning.

The proposed architecture envisions the coexistence of two complementary knowledge structures. The vector index based on the FAISS database provides fast semantic search over unstructured text of regulatory documents, while the knowledge graph encodes the domain ontology. For example, the knowledge graph may include the following types of entities and relationships:

"document" $\Longleftrightarrow$ "regulates" $\Longrightarrow$ "disaster type";
"algorithm" $\Longleftrightarrow$ "applied at phase" $\Longrightarrow$ "response phase";
"document" $\Longleftrightarrow$ "operates in jurisdiction" $\Longrightarrow$ "constituent entity of the Russian Federation";
"algorithm" $\Longleftrightarrow$ "defines action for" $\Longrightarrow$ "executor role";
"document" $\Longleftrightarrow$ "references" $\Longrightarrow$ "document".

The proposed architecture constitutes an act of self-organization in the sense that the system, without external training (without additional model fine-tuning), adapts the search space to the current context using only the metadata extracted from the query.

When processing a user query, the following sequence of operations is proposed. At the first stage, semantic parsing of the query is performed to extract not only attribute metadata but also to identify key concepts to be matched against the knowledge graph ontology. At the second stage, a query is made to the knowledge graph to identify the subset of documents that satisfy the extracted contextual constraints. At the third stage, semantic vector search, analogous to that described earlier, is performed within the identified document subset. At the final stage, the generative model receives not only relevant text fragments but also structured information about regulatory relationships from the knowledge graph, allowing it to formulate a response with explicit indication of source hierarchy and justification of their applicability.

The integration of the knowledge graph provides two fundamental advantages. First, in the presence of discrepancies between federal and regional algorithms, the system is able to explicitly indicate this circumstance and, guided by the relations of regulatory hierarchy, determine the priority source or present alternative options with appropriate caveats. Second, the knowledge graph opens up the possibility for proactive context expansion: even if a relevant fragment was not directly retrieved by vector search, it can be included in the generation context based on referential integrity relationships or thematic proximity in the ontological space.

5 Conclusion

In this work, we have proposed and validated a methodology for constructing a self-organizing knowledge base for a disaster medicine decision support system based on a hypergraph of regulatory documents and RAG technology. It has been shown that the regulatory framework forms a hierarchical hypergraph in which regional documents act as global semantic attractors, while federal documents act as local perturbations. Classical RAG, equivalent to nearest neighbor search using the pairwise cosine similarity metric, is insufficient for accounting for higher-order interactions (jurisdiction, disaster type, response phase, executor role) and systematically retrieves semantically similar but normatively incompatible fragments from different constituent entities of the Russian Federation, leading to compilative noise in the responses of the generative model.

To overcome this limitation, a two-stage self-organizing algorithm is proposed, which includes metadata extraction from the query using an LLM, hypergraph filtering based on this metadata, and subsequent semantic search on the reduced subset. The proposed mechanism breaks the symmetry between regional attractors, reducing the weight of local norms if the query does not explicitly specify a region.

However, this two-stage architecture currently remains a conceptual framework and has not been experimentally implemented or evaluated in the present study. All quantitative results reported in Section 3 (scores, distributions, entropy, and comparative performance of baseline, RAG, and agentic search) refer exclusively to the classical single-stage semantic search pipeline without metadata filtering. The proposed cascading self-organizing filter is a direction for future work, not a validated component of the reported experiments.

Thus, the main experimentally supported contributions of this paper are: (1) empirical demonstration that classical RAG systematically mixes regional regulatory fragments, (2) quantification of this effect via response entropy and expert evaluation, and (3) identification of overlapping regional attractors as the underlying geometric mechanism. The two-stage self-organizing hypergraph filter is presented as a theoretically motivated solution to this problem, requiring separate implementation and validation in subsequent research.

Further development of the methodology involves learning to rank on the hypergraph using graph neural networks for dynamic adjustment of weights between federal and regional sources, as well as the integration of a formal knowledge graph for logical inference and multi-step reasoning. The developed methodology can be extended to other subject domains with a hierarchical regulatory framework (environmental monitoring, social and legal protection, etc.) where intelligent decision support is required in the presence of conflicting information sources.

Data Availability The data presented in this study are available on request from the corresponding author.

Declarations

Conflict of interest The authors declare no conflict of interest.

Consent for publication All authors have read and agreed to the published version of the manuscript.

References

1. A.I. Kilnik, G.A. Bagaev, A.P. Maksachuk, M.N. Zamyatin, A.V. Osipov, S.S. Moskvina, The role of information in improving the responsiveness of the Russian ministry of health's disaster medicine service to emergencies. Disaster Med. **4**, 27–32 (2025)
2. B.V. Bobiy, Normative legal regulation and organizational and methodological support for the functioning of the disaster medicine service of the ministry of health of russia: Status and some ways to improve them. Disaster Med. **4**, 59–69 (2024)
3. A.V. Osipov, S.F. Goncharov, M.V. Bystrov, I.V. Gashigulina, O.V. Kakurin, A.I. Kilnik, G.A. Bagaev, A.P. Maksachuk, Organizational models of functioning of territorial centers of disaster medicine. Disaster Med. **2**, 23–32 (2025)
4. M. Abouzahra, J. Tan, The multi-level impact of clinical decision support system: A framework and a call for mixed methods evaluation. In: PACIS 2014 Proceedings, p. 224 (2014)
5. R. Noll, J. Windschmitt, E. Hofmann, N. Bergmann, J. Schaaf, Retrieval-augmented generation for medical decision-making in emergency care, in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. pp.1–7 (2025). IEEE
6. O.E. Karpov, D.A. Andrikov, M.N. Zamyatin, A.V. Osipov, A.I. Kilnik, G.A. Bagaev, A.P. Maksachuk, A.E. Hramov, Prospects and limitations of artificial intelligence technologies in the decision support system of the disaster medicine service. Med. Doctor Inform. Technol. **4**, 6–15 (2025)
7. X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, Searching for best practices in retrieval-augmented generation. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17716–17736 (2024)
8. Y. Miao, Y. Zhao, Y. Luo, H. Wang, Y. Wu, Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: Scoping review. J. Med. Internet Res. **27**, 80557 (2025)
9. M. Abo El-Enen, S. Saad, T. Nazmy, A survey on retrieval-augmentation generation (rag) models for healthcare applications. Neural Comput. Appl. **37**(33), 28191–28267 (2025)
10. S.A. Beber, K.D. Groff, T.R. Mange, J.T. Bram, P.D. Fabricant, Not ready for prime time: Limitations of a retrieval-augmented generation large language model in assessing risk of bias in observational studies. J. Pediatr. Soc. N. Am. **14**, 100294 (2025). https://doi.org/10.1016/j.jposna.2025.100294
11. H.S. Wong, T.K. Wong, Multi-evidence clinical reasoning with retrieval-augmented generation for emergency triage: Retrospective evaluation study. JMIR Med. Inform. **14**(1), 82026 (2026)
12. S. Liu, A.P. Wright, A.B. McCoy, S.S. Huang, B. Steitz, A. Wright, Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. J. Am. Med. Inform. Assoc. **32**(6), 1032–1039 (2025)
13. V.S. Khorev, S.A. Kurkin, G. Zlateva, R. Paunova, S. Kandilarova, M. Maes, D. Stoyanov, A.E. Hramov, Disruptions in segregation mechanisms in fmri-based brain functional network predict the major depressive disorder condition. Chaos, Solitons Fractals **188**, 115566 (2024)

14. A.V. Andreev, V.A. Maksimenko, A.N. Pisarchik, A.E. Hramov, Synchronization of interacted spiking neuronal networks with inhibitory coupling. Chaos, Solitons Fractals **146**, 110812 (2021)
15. Y. Xiong, Y. Chen, H. Zhang, ICR: A framework for resolving knowledge conflicts in retrieval-augmented generation. Neurocomputing **664**, 132139 (2025)
16. M. Yazaki, S. Maki, T. Furuya, K. Inoue, K. Nagai, Y. Nagashima, J. Maruyama, Y. Toki, K. Kitagawa, S. Iwata, Emergency patient triage improvement through a retrieval-augmented generation enhanced large-scale language model. Prehosp. Emerg. Care **29**(3), 203–209 (2025)
17. C. Merola, J. Singh, Reconstructing context: evaluating advanced chunking strategies for retrieval-augmented generation. Int. Workshop Knowl. Enhanced Inform. Retrieval (Cham: Springer Nature Switzerland, 2025), pp. 3–18
18. S.F.C. Haviana, M.A. Riyadi, R. Kusumaningrum, Evaluation of chunking strategies in rag application for explicit retrieval on indonesian language scientific papers, in *2025 12th International Conference on Electrical Engineering* Computer Science and Informatics (EECSI). IEEE (2025), pp.59–65
19. S. Myers, T.A. Miller, Y. Gao, M.M. Churpek, A. Mayampurath, D. Dligach, M. Afshar, Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. J. Am. Med. Inform. Assoc. **32**(2), 357–364 (2025)
20. H. Jégou, M. Douze, J. Johnson, L. Hosseini, C. Deng, Faiss: similarity search and clustering of dense vectors library. Astrophysics Source Code Library, 2210 (2022). https://ui.adsabs.harvard.edu/abs/2022ascl.soft10024J/abstract
21. Y. Tao, A. Hiatt, E. Haake, A.J. Jetter, A. Agrawal, When context leads but parametric memory follows in large language models, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (2024), pp. 4034–4058
22. Z. Wang, Y. Li, J. Yan, Y. Cheng, Y. Zhang, Unveiling attractor cycles in large language models: a dynamical systems view of successive paraphrasing. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by W. Che, J. Nabende, E. Shutova, M.T. Pilehvar, Association for Computational Linguistics, Vienna, Austria (2025), pp. 12740–12755. https://doi.org/10.18653/v1/2025.acl-long.624
23. M. Perełkiewicz, S. Dadas, R. Powiata, Smclm: Semantically meaningful causal language modeling for autoregressive paraphrase generation. IEEE Access **13**, 119197–119214 (2025). https://doi.org/10.1109/ACCESS.2025.3585679
24. S. Farquhar, J. Kossen, L. Kuhn, Y. Gal, Detecting hallucinations in large language models using semantic entropy. Nature **630**(8017), 625–630 (2024)
25. L. Deng, H. Hu, K. Lu, P. He, Llm-augmented multi-agent cooperative framework for medical case retrieval in cardiology. J. King Saud Univ. Comp. Inform. Sci. **37**(9), 267 (2025)
26. M.S. Baysan, S. Uysal, İ İşlek, Ç. Çığ Karaman, T. Güngör, Llm-as-a-judge: automated evaluation of search query parsing using large language models. Front. Big Data **8**, 1611389 (2025)

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.