ISSN 2007-9737

Entropy–Distance Approach to Evaluating Diversity and Robustness in Organizational Information Retrieval

Shakarim Aubakirov[1], Iskander Akhmetov[1,2], Alexander Krassovitsky[2], Alexander Gelbukh[3,*]

[1] Kazakh-British Technical University,
Kazakhstan

[2] Institute of Information and Computational Technologies,
Kazakhstan

[3] Instituto Politécnico Nacional, CIC,
Mexico

a.krassovitskiy@iict.kz, {sh_aubakirov,i.akhmetov}@kbtu.kz,
gelbukh@cic.ipn.mx

Abstract. Information retrieval constitutes a critical component of organizational information management, directly affecting the efficiency, accuracy, and resilience of decision-making processes. Conventional evaluation metrics—such as precision or click-through rates—do not adequately capture the lexical and semantic diversity of retrieved content, limiting their utility in managerial contexts where both relevance and variety are essential. This study introduces a scalable, language-agnostic entropy–distance framework designed to assess the robustness of retrieval systems under controlled linguistic variation. The framework integrates Shannon entropy, to quantify lexical diversity, with semantic dispersion measures derived from SBERT embeddings, enabling joint evaluation of breadth and coherence in search outputs. Using a curated 6.6M-article Wikipedia corpus, topics were clustered, summarized, and reformulated into paraphrased queries, which were executed across Google, Bing, and DuckDuckGo. The resulting outputs reveal significant differences in diversity–coherence trade-offs between platforms, with DuckDuckGo exhibiting the highest adaptability to query variation. The proposed methodology supports information governance by providing an unsupervised, reproducible metric that enables comparative auditing of search performance in enterprise and public domains. The findings offer actionable insights for optimizing retrieval strategies, mitigating systemic bias, and enhancing the resilience of organizational search infrastructures.

Keywords. Entropy, search engine evaluation, paraphrase robustness, semantic dispersion, information retrieval, query variability.

1 Introduction

Search engines and information retrieval platforms constitute integral components of contemporary *information management* infrastructures, enabling the efficient location, access, and exploitation of knowledge resources. Embedded in enterprise search systems, corporate knowledge bases, and business intelligence (BI) dashboards, these tools operate as critical enablers within the *information management lifecycle*, encompassing discovery, storage, retrieval, and utilization of information for organizational decision-making [14].

The evaluation of retrieval output quality in such systems remains a complex and unresolved challenge. Official documentation by Google acknowledges that the assessment of search result quality is inherently multifaceted, requiring human raters and behavioral analytics [22]. Traditional Search Engine Optimization (SEO) indicators—such as keyword frequency, backlink counts, or Google's PageRank—previously functioned as proxies for retrieval quality but

doi: 10.13053/CyS-29-4-5824

now represent narrow, web-centric measures that are insufficient for context-driven, enterprise-scale search environments [24].

The increasing integration of semantic interpretation and user intent [28] necessitates the adoption of evaluation frameworks extending beyond keyword- and link-centric approaches to meet organizational requirements for reliable, diverse, and relevant access to information.

A persistent challenge in both public and corporate search systems involves the *management of diversity* within retrieved content. In enterprise contexts, insufficient diversity may manifest as the repeated retrieval of identical policy documents or duplicated records from a single departmental repository, thereby constraining exposure to alternative perspectives. Such homogeneity may indicate systemic bias or ranking inefficiencies, concentrating informational outputs within a limited set of sources. In the public domain, Google addressed this issue through the “site diversity” update, which restricts most queries to no more than two top results from the same domain [21]; analogous phenomena are observable within corporate intranets or BI environments when a dominant repository monopolizes output rankings. Conversely, excessive variability—where top-ranking items are loosely related or tangential to the query—can reduce relevance and hinder operational decision-making. Comparable effects have been noted in broad e-commerce search contexts [19].

The consequences of such imbalance are particularly significant in organizational settings, where timely and accurate access to relevant yet diverse information supports effective decision-making, risk assessment, and innovation. Overly homogeneous outputs foster informational silos, while excessive heterogeneity introduces noise and decision inertia. Standard Information Retrieval (IR) evaluation metrics, including precision and Normalized Discounted Cumulative Gain (NDCG), fail to account for these nuances. Two retrieval sets can obtain equivalent relevance scores while differing substantially—one consisting of repeated instances of the same document, the other providing distinct perspectives from multiple organizational units [19]. Consequently, diversity constitutes a strategic dimension of *information quality* that remains insufficiently captured by conventional evaluation measures.

This study incorporates **Shannon entropy** as a quantitative indicator for assessing diversity (or uncertainty) in retrieval outputs [27]. Commonly applied in ecology to measure biodiversity, entropy is here adapted to assess the distributional spread of information sources or topical categories within retrieval results. In organizational contexts, entropy serves as a diagnostic indicator of an information system’s operational health, enabling the detection of over-concentrated or sufficiently varied output distributions. For example, if all top-ranked items originate from the same department or repository, entropy approaches zero; in contrast, balanced representation across multiple relevant units increases entropy [25]. While previous studies have examined entropy within the context of search diversity [35], its deployment as a management-oriented evaluation tool for corporate search platforms remains underdeveloped. This situates entropy within a broader *information diversity management* agenda, linking it to key performance indicators (KPIs) for organizational information systems.

The underlying hypothesis is that entropy in retrieval outputs correlates with the degree of thematic coherence or divergence across search systems responding to an identical informational requirement. High entropy may reflect interpretive ambiguity or multifaceted query intent, whereas low entropy may indicate authoritative convergence. Diversity alone, however, is not inherently advantageous unless balanced with relevance. A high-entropy output may broaden informational perspectives but risks incoherence if relevance is compromised. For this reason, entropy is combined with a *semantic dispersion* metric to capture both lexical variety and conceptual breadth, thus enabling the assessment of an organization’s capability to maintain coherence while facilitating access to diverse knowledge resources.

To operationalize this approach, a unified **entropy–distance score** is proposed, jointly capturing lexical diversity and semantic dispersion within retrieval results. The metric is evaluated across three major search platforms (Google, Bing, and DuckDuckGo) using a controlled set of paraphrased queries. While the empirical evaluation employs public search engines, the framework is directly transferable to enterprise search and knowledge retrieval environments, providing a scalable, language-agnostic, and unsupervised diagnostic for auditing retrieval robustness under varied query formulations. This methodology supports *information governance* objectives by delivering measurable insights into the extent to which retrieval systems preserve diversity and relevance in dynamic information environments.

The remainder of this paper is structured as follows. section 2 reviews related work on retrieval evaluation, diversity metrics, and information-theoretic measures within information systems. section 3 details the data acquisition process and analytical methodology, including entropy and semantic dispersion computation. section 4 presents empirical findings, highlighting diversity patterns and inter-system differences. section 5 discusses implications for organizational information management, outlines study limitations, and suggests avenues for further research. Finally, section 6 concludes with practical recommendations for embedding diversity-oriented evaluation mechanisms into information management practices.

2 Literature Review

Entropy, as originally formalized in information theory by Shannon [23], has evolved into a versatile analytical construct across multiple domains of information systems. Its capacity to quantify **uncertainty**, **diversity**, **disagreement**, and **ambiguity** has made it relevant not only for search and recommendation systems, but also for summarization, question answering (QA), and reasoning with large language models (LLMs) in enterprise-scale environments. Within organisational information systems and knowledge management infrastructures, these capabilities are critical for assessing the reliability, transparency, and robustness of outputs generated by complex, user-facing, and often generative systems. This section presents a critical, thematically structured review of prior work across five interrelated domains—search systems, recommender systems, summarisation, QA, and LLM-based reasoning—focusing explicitly on methodological limitations, conceptual inconsistencies, and unresolved challenges that motivate the proposed cross-system entropy-based evaluation framework.

These domains are not isolated silos; rather, they form an interconnected pipeline in which search systems retrieve candidate content, recommender and knowledge-based systems prioritise and filter it, summarisation condenses it, and QA or LLM-based reasoning modules synthesise it into actionable knowledge. Entropy, in this context, can act as a unifying metric traversing all these stages.

Entropy and Search Systems

Evaluation of search systems in both public web and enterprise contexts has traditionally relied on surface-level behavioural indicators, such as click-through rate (CTR) and dwell time. While these metrics are operationally convenient, they fail to account for the **semantic diversity** and **ambiguity** that often characterise retrieved document sets. In enterprise environments, where queries may be underspecified, domain-specific, or long-form [13], such omission leads to an incomplete assessment of retrieval quality.

An alternative approach employs *click entropy*—a measure of distributional uncertainty in user interaction patterns—to capture query ambiguity and informational richness. For instance, [4] propose an entropy-biased framework that prioritises low-frequency, high-specificity clicks, operationalised via an *inverse query frequency* (IQF) measure to enhance query representation in click graphs. More recent enterprise-focused studies, such as [34] and [32], extend this notion beyond user logs, incorporating entropy-based assessments into the evaluation of retrieval-augmented generation (RAG) pipelines and domain-specific QA systems.

Although diversity-aware ranking has been implemented in isolated industrial settings—e.g., Google's site diversity adjustments [21]—optimisation strategies largely remain heuristic or rule-based. Moreover, contemporary audits of search optimisation practices [14, 24] indicate continued dependence on outdated proxies such as link volume and keyword density, which are agnostic to semantic redundancy.

Comprehensive surveys [29] emphasise that in both search and recommendation domains, over-optimisation on homogenous training corpora can lead to a *quality saturation effect* [17], wherein retrieval accuracy plateaus despite increased data volumes. In such conditions, entropy-based evaluation provides a mathematically grounded alternative for quantifying semantic diversity, novelty, and system robustness—capabilities that remain underexplored in cross-system or cross-domain benchmarking.

While search systems primarily address the retrieval stage, enterprise environments often require subsequent filtering and personalization, where recommender and knowledge-based systems play a central role. The entropy-focused evaluation paradigm thus naturally extends from search into recommendation.

Entropy in Recommendation and Knowledge-Based Systems

In enterprise environments, recommender systems extend beyond product suggestions to encompass document retrieval, expert identification, and project matchmaking. In such contexts, entropy serves as a quantitative indicator of **information diversity**, **preference uncertainty**, and **semantic coverage**. For example, Lee [12] incorporates entropy into collaborative filtering to capture global rating variance, enabling systems to detect unstable or weakly correlated preferences. Similarly, Yalcin et al. [31] apply entropy to group recommendations, identifying polarizing resources whose relevance varies significantly across organizational subgroups.

Recent work in knowledge-intensive settings highlights the role of entropy in maintaining balanced retrieval coverage. Jiang et al. [6] and Li et al. [13] show that integrating diversity-oriented retrieval—often measured indirectly via entropy—improves the match between long-form, semantically rich queries and domain-specific knowledge bases. In industrial semantic search, Naqvi et al. [16] demonstrate that entropy-based signals help surface underrepresented maintenance insights, avoiding over-optimization toward redundant or frequently accessed documents.

Entropy has also been embedded in reinforcement learning-based recommenders to balance *novelty* and *relevance* [3] and in topic extraction modules [7] to control the dispersion of latent semantic factors. However, despite these advances, existing approaches predominantly optimize entropy *within* a single recommender or retrieval pipeline. No comparative frameworks currently assess entropy across multiple enterprise knowledge systems or evaluate its interaction with semantic distance measures—gaps that the present study directly addresses in the context of cross-engine search result evaluation.

However, the output of recommendation and knowledge-based retrieval often feeds directly into summarisation pipelines, especially in organisational decision-making contexts. This creates an additional layer where entropy can assess whether the diversity and uncertainty present in upstream retrieval are preserved or lost during information condensation.

Entropy in Summarization and Enterprise Knowledge Distillation

In organizational contexts, summarization is not limited to news or open-domain text but is integral to condensing large volumes of internal reports, meeting transcripts, and technical documentation into decision-ready insights. In such workflows, entropy provides a formal mechanism for assessing **semantic coverage** and **content diversity**, ensuring that summaries do not omit critical but less frequent informational elements.

In extractive summarization, Khurana et al. [9] introduce E-Summ, which applies Non-negative Matrix Factorization (NMF) to detect latent semantic units and computes the entropy of their distributions to score sentences. Higher entropy correlates with broader topic coverage, offering a transparent alternative to purely heuristic or reference-based evaluation methods. This is particularly relevant for enterprise knowledge distillation, where coverage of minority but high-value topics can be critical for risk assessments or strategic planning.

In abstractive summarization, Xu et al. [30] analyze token-level entropy during decoding in models such as BART and PEGASUS, finding that low-entropy tokens correspond to extractive copying while high-entropy tokens align with generative novelty. This distinction is relevant for monitoring the balance between innovation and factual reliability in automated summarization of corporate knowledge bases.

Despite these advances, most entropy-based approaches in summarization focus exclusively on the generated text, overlooking the upstream quality of the retrieved content. In enterprise environments, summaries are often derived from heterogeneous search outputs or multi-source knowledge retrieval. The present study addresses this gap by evaluating entropy not at the summary layer but at the *search result layer*, enabling quality control of the input space before summarization is performed. This approach strengthens governance over downstream knowledge synthesis and aligns with information management quality frameworks.

Summarisation outputs frequently serve as inputs for question answering systems, whether in open-domain or enterprise settings. Consequently, entropy-based analysis at the QA stage can reveal whether ambiguity and uncertainty propagate, amplify, or diminish as information passes through increasingly interpretive stages of processing.

Ambiguity and Entropy in Enterprise QA Systems

In enterprise environments, question answering (QA) systems are increasingly integrated with corporate knowledge bases, internal documentation, and structured repositories to support decision-making. Ambiguity in such contexts may arise from inconsistent terminology, overlapping data sources, or incomplete metadata. Entropy provides a mathematically grounded means of quantifying this uncertainty, enabling quality control in both retrieval and answer synthesis stages.

While open-domain QA benchmarks such as *AmbigQA* [15] and its extensions [10] formalize the concept of multiple plausible answers, these frameworks rarely address organizational requirements such as auditability, compliance, or governance over answer generation. More recent work has introduced entropy explicitly as a proxy for semantic uncertainty. For example, in medical QA, Wang et al. [26] compute word-sequence entropy to detect unstable answers, while Kuhn et al. [11] apply semantic entropy for hallucination detection.

Such approaches are relevant to enterprise QA pipelines, where incorrect or inconsistent answers can lead to operational risks. Unlike model-specific confidence scores, entropy is model-agnostic and can be applied uniformly across heterogeneous QA subsystems, making it suitable for enterprise-wide monitoring of knowledge access reliability.

Modern QA pipelines, particularly those augmented with large language models, blur the line between answering and reasoning. This makes it necessary to examine how entropy functions not only as a measure of retrieval or answer ambiguity, but also as a guide for multi-step reasoning and generative synthesis in LLM-driven systems.

Entropy in LLM Reasoning and Enterprise Knowledge Generation

Large Language Models (LLMs) are increasingly deployed as reasoning engines within enterprise search and decision-support systems, where they must synthesize multi-source data into coherent narratives or recommendations. Entropy-based methods have been adopted for selecting prompts, guiding reasoning paths, and filtering unreliable outputs. For instance, the INFORM framework [33] uses entropy to select and generate multi-step reasoning prompts. In hallucination detection, Farquhar et al. [5] propose *semantic entropy* as a robust signal for identifying fabricated content, a capability critical in regulated industries.

Further developments, such as [1] and [20], focus on entropy minimization during decoding to improve coherence and factual consistency. Other works propose entropy-based sentence-level hallucination scores [8] and fine-grained semantic entropy estimators [18], moving beyond token-level probability into latent semantic evaluation.

In enterprise contexts, these methods can be embedded in Retrieval-Augmented Generation (RAG) pipelines to ensure that generated content remains consistent with verified corporate data sources. By incorporating entropy as a governance metric, organizations can track the stability of LLM outputs over time, detect shifts in semantic alignment with internal knowledge bases, and prevent knowledge drift in long-term deployments.

In enterprise deployments, where accuracy, auditability, and compliance are critical, such reasoning capabilities are integrated into complex multi-source knowledge systems. This necessitates a refined entropy-based governance framework that operates across interconnected subsystems, from retrieval to reasoning, within a single organisational ecosystem.

Research Gap

Across the reviewed domains, entropy consistently emerges as a robust, **model-agnostic** indicator for **ambiguity**, **disagreement**, and **uncertainty**. Nevertheless, prior studies have primarily focused on isolated systems—such as a single question answering model, a recommender platform, or a standalone search engine—without extending the analysis to multi-system environments typical of enterprise information infrastructures.

No prior research has systematically examined entropy distributions across *multiple knowledge retrieval platforms*, including corporate search systems, federated search environments, or hybrid enterprise–public search architectures. Furthermore, the literature does not address the *combined use of entropy and semantic distance* as a stability and coherence indicator in knowledge delivery pipelines.

The present work addresses this gap by (i) computing entropy for outputs generated by multiple search platforms when processing paraphrased queries, (ii) comparing entropy values with inter-system semantic divergence, and (iii) interpreting entropy–distance patterns as indicators of systemic disagreement, bias concentration, or instability. This comparative, cross-system perspective extends entropy analysis from a theoretical diagnostic to a practical **governance metric** for assessing the **interpretability**, **reliability**, and **trustworthiness** of enterprise information access systems.

Across this interconnected workflow—from initial retrieval to final reasoning—entropy emerges as a common analytical thread. Yet, despite its applicability at each stage, the literature treats these domains separately, missing the opportunity for cross-system, end-to-end evaluation.

3 Methodology

This section presents the full methodological pipeline for evaluating entropy-based diversity in search engine results. The process is organized into three components. The methodology begins with the acquisition and preprocessing of a large-scale textual corpus extracted from Wikipedia. This is followed by a structured overview of the analytical workflow, including the construction of semantic representations, their clustering, summarization, and transformation into query formulations for entropy-based evaluation. Finally, the implementation details are outlined, covering vectorizer benchmarking, query execution, and the operational setup used to compute entropy and semantic distance metrics across multiple search engines.

3.1 Data Acquisition and Preprocessing

This study relies on large-scale textual data extracted from the full English-language Wikipedia corpus. The foundational dataset was obtained by downloading the compressed dump `enwiki-20250601-pages-articles-multistream.xml.bz2` from the Wikimedia data archive[1]. The file contains the complete set of Wikipedia articles as of June 1, 2025, totaling approximately 109 GB in compressed format.

The raw dump was parsed and cleaned using `WikiExtractor`, a standard open-source tool designed to remove MediaWiki markup and output plain-text articles. To accelerate processing, extraction was performed in parallel using 8

[1] `https://meta.wikimedia.org/wiki/Data_dump_torrents#English_Wikipedia`

CPU cores, yielding an average throughput of roughly 4,220 articles per second. The full extraction process completed in 4,339 seconds (approximately 1 hour and 12 minutes), producing 18,315,148 cleaned articles stored in JSON format, distributed across 18 separate output files.

Following extraction, a filtering stage was applied to remove empty or trivially short documents. Specifically, articles with fewer than 50 non-whitespace characters were excluded. This threshold ensured semantic substance and structural completeness in the retained dataset. The filtered corpus comprises 6,649,690 articles and this corresponds to a validity ratio of 36.31%—i.e., the proportion of non-trivial articles retained after filtering. The breakdown is summarized below:

- Total extracted articles: 18,315,147
- Short or empty articles: 11,665,457
- Retained (clean) articles: 6,649,690
- Validity ratio: 36.31%

Each retained article was stored as a separate JSON object with minimal metadata (e.g., title and content), facilitating downstream processing in vectorization, clustering, and summarization stages. All subsequent steps in the methodology—including topic modeling, query generation, and entropy analysis—were based exclusively on this curated subset of Wikipedia content.

3.2 Workflow Overview

The analytical workflow comprises multiple stages, each designed to transform raw Wikipedia-derived topics into measurable search result diversity indicators. The rationale for each stage is as follows.

Text Vectorization transforms raw text into dense numerical representations, enabling clustering in continuous semantic space and forming the basis for similarity-based operations.

Clustering of Articles reveals latent topical structures by grouping semantically similar articles. This enables controlled topic-level analysis.

Summarization compresses each article into a concise extractive representation, improving interpretability and facilitating downstream operations such as abstraction and labeling.

Cluster-Level Abstraction combines article summaries within each cluster and re-summarizes them to obtain a compressed thematic synopsis—representing the central idea of the topic cluster.

Topic Labeling via LLMs uses GPT-4.1[2] to assign semantic labels to cluster summaries. These labels guide subsequent query formulation.

Query Paraphrasing ensures controlled linguistic variation and simulates realistic user behavior in query formulation. Ten paraphrasing strategies were applied per topic label, each reflecting a different linguistic transformation pattern (e.g., synonym replacement, clarification, style shift, code-switching).

Search Engine Querying provides access to real-world result sets for each paraphrased query, retrieved independently from three commercial search engines.

Entropy Computation measures lexical unpredictability in search result snippets. Higher entropy implies higher diversity in language use, while lower entropy suggests convergence or redundancy.

Semantic Distance Estimation (via cosine similarity) quantifies semantic cohesion across search results. Smaller cosine distances between snippet vectors indicate tighter topical focus and higher semantic overlap.

Comparative Evaluation assesses the behavior of search engines by comparing entropy and semantic distance across query types, topics, and platforms. This allows entropy to serve as a proxy for systemic disagreement or manipulation tendencies.

To unify these metrics into a single diagnostic signal, entropy values were normalized by their corresponding cosine distances. The resulting *entropy-to-distance ratio* amplifies high lexical variability when paired with semantic inconsistency—thus capturing not only diversity

[2] GPT-4.1 is an advanced version of OpenAI's large language model, capable of multi-turn reasoning, summarization, and semantic labeling tasks.

but potential topical incoherence or noise in search outputs.

3.3 Experiment Design

The dataset consists of 6,649,690 cleaned Wikipedia articles extracted from a compressed XML dump using the WikiExtractor tool. Articles shorter than 50 non-whitespace characters were filtered out to ensure semantic substance. All articles were stored in JSONL format with minimal metadata.

To determine the most suitable vectorization technique for downstream semantic clustering and analysis, a benchmark was conducted on a stratified classification task. From the preprocessed corpus of 6,649,690 Wikipedia articles, a filtered subset of 487,424 articles was isolated based on the presence of explicit topical markers in their category metadata. The filtered topics were limited to ten high-level domains: *economics, sociology, mathematics, physics, chemistry, biology, medicine, music, history*, and *politics*.

This topic set was chosen deliberately. An initial analysis of the full corpus revealed that a significant portion of the articles consisted of structurally repetitive templates—such as annual birth/death lists and biographical stubs—which lacked sufficient semantic depth for robust vectorization or clustering. The selected ten domains instead reflect meaningful areas of human knowledge and activity, providing topical diversity and conceptual substance for evaluation purposes.

To evaluate vectorization performance reliably, the optimal sample size was computed based on standard statistical assumptions: 95% confidence level, 1% margin of error, and maximal variance. This yielded a lower bound of 9,604 samples. Applying finite population correction, the sample size adjusted to 9,419. To enforce topic balance and simplify stratification, a uniform sample of 1,000 articles per domain was extracted, resulting in a benchmark set of 10,000 articles.

Four vectorization methods were tested—*TF-IDF*, *FastText*, *SBERT*, and *E5*. For each method, the 10,000 samples were embedded and used as input to a logistic regression classifier. Performance was evaluated via accuracy, macro/micro F1 scores, and inference time. Results are summarized in Table 1.

While E5 embeddings demonstrated the highest classification performance, their inference time (approximately 8.8 hours for full-scale encoding) rendered them infeasible for processing millions of documents. FastText and TF-IDF were faster but underperformed in accuracy and semantic fidelity. SBERT offered the best tradeoff between semantic quality and computational efficiency (70 minutes total for 6M articles), making it the most practical candidate for high-volume semantic processing.

Consequently, SBERT was selected as the default vectorization method for all downstream operations, including clustering, semantic similarity computation, and entropy-based analysis.

Clustering was implemented using MiniBatch K-Means (scalable for large data) and HDBSCAN (density-based, noise-aware). MiniBatch K-Means was selected based on runtime feasibility and topic interpretability.

Summarization relied on the GreedSum algorithm [2], which applies greedy optimization and variable neighborhood search to select maximally informative, non-redundant sentences from article text[3].

Each cluster summary was labeled using GPT-4.1. For each resulting topic label, ten query variants were generated using a locally hosted LLaMA3 model exposed via HTTP API. Prompts specified linguistic transformation types to ensure paraphrastic diversity.

Query execution was conducted on Google (via official API), Bing, and DuckDuckGo (via browser automation agents). Yandex and Yahoo were excluded due to excessive runtime (exceeding 25 hours per batch).

Returned search snippets were preprocessed to remove HTML artifacts, tokenized at word level, and processed using entropy and cosine similarity metrics.

Comparative evaluation was performed along two axes: (i) engine-wise and (ii) paraphrase-wise. This enabled empirical assessment of entropy

[3] `https://github.com/iskander-akhmetov/Greedy-Summarization`

Table 1. Comparison of vectorizers based on classification metrics and inference time. Efficiency is computed as Accuracy / Time_sec

Vectorizer	Accuracy	F1_macro	F1_micro	Time_sec	Efficiency
E5	0.8475	0.8474	0.8475	53	0.016
SBERT	0.8240	0.8240	0.8240	7	0.118
FastText	0.6505	0.6452	0.6505	15	0.043
TF-IDF	0.8360	0.8367	0.8360	12	0.070

patterns and semantic divergence across search platforms. The complete methodological pipeline is summarized in Figure 1.

4 Results

This section reports empirical findings for each major component of the pipeline. Emphasis is placed on vectorizer performance, clustering scalability, paraphrasing logic, entropy estimation, and semantic coherence of search results. Each subsection corresponds to a specific stage of analysis. Comparative evaluations are presented across vectorization methods, cluster structures, query variations, and search engine outputs.

4.1 Clustering Strategy

From the full corpus of 6,506,091 Wikipedia articles, a random sample of 500,000 articles was selected to reduce computational overhead. Although no neural networks were involved in this phase, this subsample was sufficient for structural exploration of category distributions and semantic grouping. Within the original corpus, 1,764,606 unique article categories were identified. The 500K subset retained 761,619 unique categories—capturing 43.16% of the total—providing a representative and diverse sample.

Clustering was benchmarked using two unsupervised algorithms: *MiniBatch K-Means* and *HDBSCAN*. Despite its density-based flexibility, HDBSCAN failed to complete execution after two hours. In contrast, MiniBatch K-Means finished in under a minute on the same data. Given the magnitude of the corpus, MiniBatch K-Means was selected as the pragmatic choice due to its scalability and efficiency.

To determine the optimal number of clusters, several internal evaluation metrics were computed across a range of cluster counts. These included the Calinski–Harabasz Index (Figure 2), Davies–Bouldin Index (Figure 3), Silhouette Coefficient (Figure 4), and Elbow Method (Figure 5). Although no clear maximum was observed across all metrics, a compromise point was identified at **15 clusters**, which balanced intra-cluster compactness and inter-cluster separation across multiple scores.

Once the number of clusters was fixed at 15, the corresponding cluster centroids were extracted and subsequently used for topic summarization and query generation.

Descriptive statistics for the clustered corpus are presented in Table 2. The table shows the number of documents originally assigned to each of the 15 clusters, as well as the average length of these documents in words and sentences prior to summarization. The clusters vary substantially in both size and verbosity: for example, Cluster 2 includes shorter texts (on average 156 words and 10 sentences), while Cluster 12 aggregates the longest documents (averaging over 700 words and 36 sentences). These disparities highlight the semantic and structural heterogeneity present in the source data across different clusters.

4.2 Article Summarization and Medoid Selection

Following the clustering of 500,000 articles using the MiniBatchKMeans algorithm, an extractive summarization step was applied to compress each article into its most informative content. The summarization procedure was implemented using the GreedySum algorithm [2], which employs a

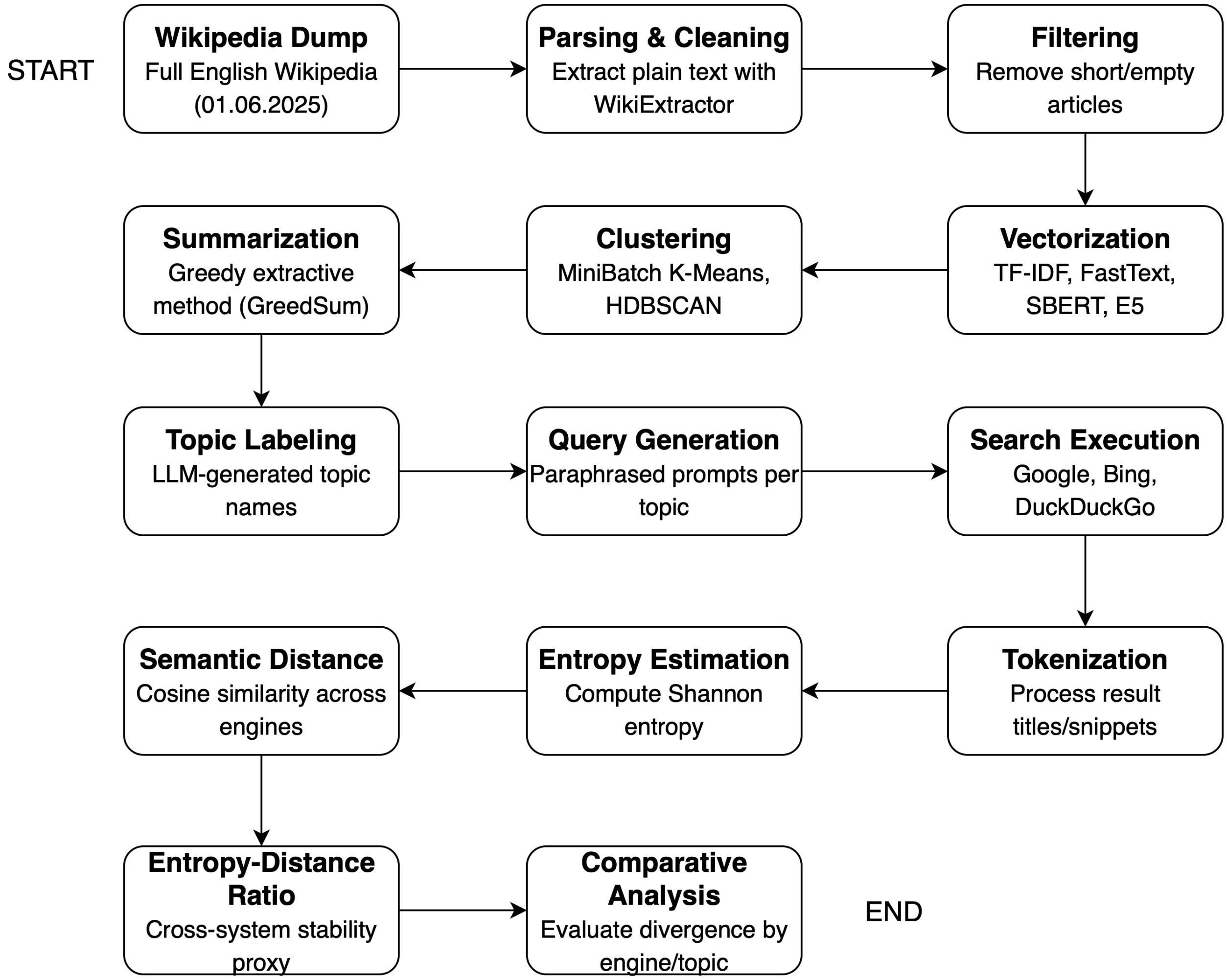
START
Wikipedia Dump
Full English Wikipedia (01.06.2025)
Parsing & Cleaning
Extract plain text with WikiExtractor
Filtering
Remove short/empty articles
Vectorization
TF-IDF, FastText, SBERT, E5
Clustering
MiniBatch K-Means, HDBSCAN
Summarization
Greedy extractive method (GreedSum)
Topic Labeling
LLM-generated topic names
Query Generation
Paraphrased prompts per topic
Search Execution
Google, Bing, DuckDuckGo
Tokenization
Process result titles/snippets
Entropy Estimation
Compute Shannon entropy
Semantic Distance
Cosine similarity across engines
Entropy-Distance Ratio
Cross-system stability proxy
Comparative Analysis
Evaluate divergence by engine/topic
END

Fig. 1. Full pipeline of the methodology, from data extraction to comparative analysis

variable neighborhood search strategy to select a diverse and representative set of sentences from scientific or encyclopedic texts. This process yielded compressed versions of all articles, enabling efficient subsequent semantic operations.

After the initial summarization of all 500,000 clustered articles using the GreedySum algorithm, a second-level summarization step was applied to abstract high-level themes. For each of the 15 clusters, individual summaries were concatenated and re-summarized into a concise 10-sentence representation capturing the dominant semantic content of the cluster. These compressed representations were then submitted to GPT-4.1 with a prompt to generate human-readable topic labels and thematic descriptions.

The generated cluster labels and descriptions are listed below:

— **Cluster 0 — Acting Debuts and Regional Demographics**: Biographical accounts of debuting actors, demographic profiles of small towns, administrative divisions in China and Iran, and historical regional reforms.

— **Cluster 1 — Political Careers and Governmental Reforms**: Ministerial posts, failed

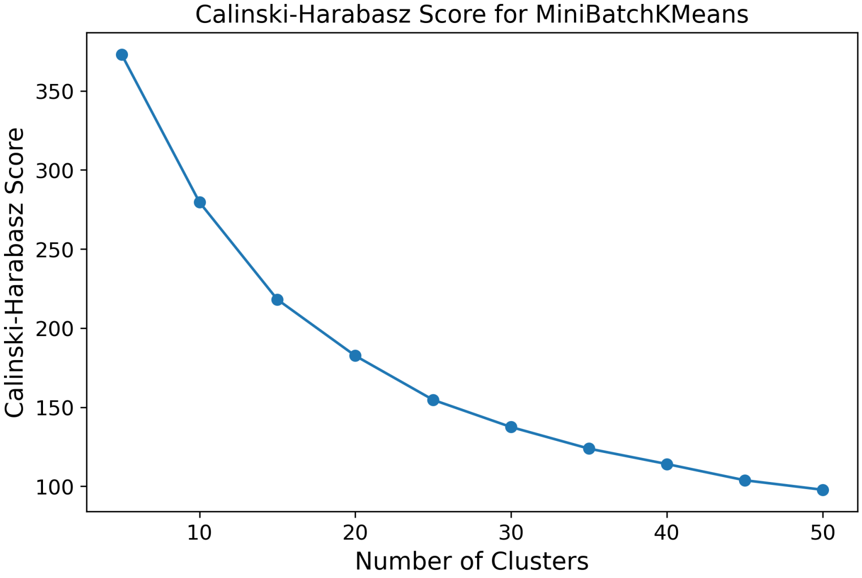

Fig. 2. Calinski–Harabasz Index for different cluster counts using MiniBatch K-Means

re-elections, ambassadorial appointments, political prisoners, electoral redistricting, and legislative coalitions.

- **Cluster 2 — Taxonomy and Species Migration**: Descriptions of diatoms, invasive plant distribution, extinct animal lineages, marine biodiversity patterns, and global species migration.
- **Cluster 3 — Musical Groups and Discographies**: Artist discographies, album re-releases, chart histories, music label involvement, and lyrical themes.
- **Cluster 4 — Military History and Colonial Campaigns**: Colonial wars, tactical reports, local governance transitions, religious proclamations, sieges, and geopolitical mapping.
- **Cluster 5 — Sports Transfers and Club Trajectories**: Player biographies, contract histories, media reporting on football clubs, and regional tournament participation.
- **Cluster 6 — Literary Diaries and Artistic Studies**: Stolen diaries, presidential biographies, literary criticism, psychological portraits, and academic correspondence.
- **Cluster 7 — Evolution of Professional Tours**: Restructuring of professional leagues, Olympic performances, coaching records, sports statistics, and global tournaments.
- **Cluster 8 — Administrative Regions and Infrastructure**: Territorial divisions, district reforms, communal governance, architectural structures, and urban planning.
- **Cluster 9 — Academic and Sports Competitions**: Participation in rugby and basketball championships, football records, university matches, and sports biographies.
- **Cluster 10 — Palace Structures and Media Industry**: Noble hierarchies, radio shows, award ceremonies, drama debuts, festival circuits, and theatrical adaptations.
- **Cluster 11 — Legislative Transformations and Elite Politics**: Redistricting reforms, petitions, agricultural modernization, budget crises, and elite educational careers.
- **Cluster 12 — Technical Specifications and Curricular Programs**: Engine configurations,

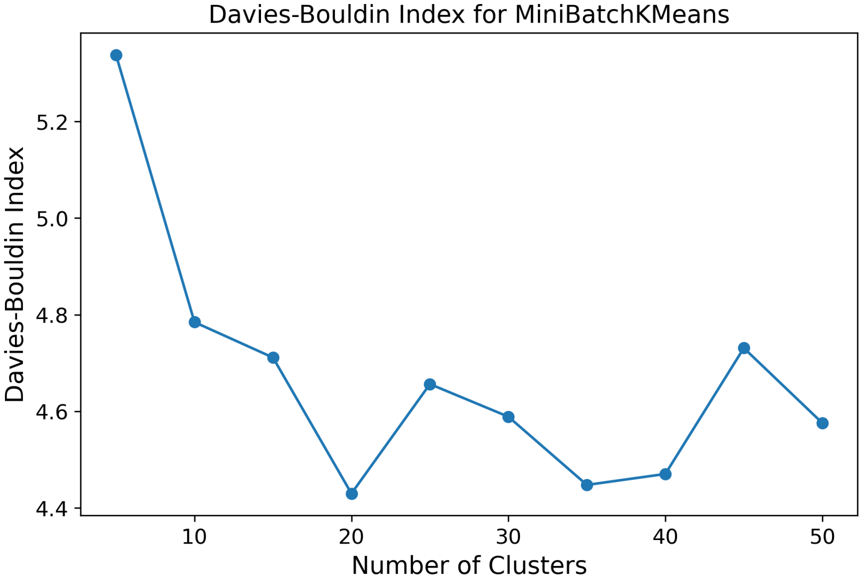

Fig. 3. Davies–Bouldin Index for different cluster counts using MiniBatch K-Means

aircraft production shutdowns, user experience design, transit systems, and medical course outlines.

— **Cluster 13 — Migrations and Political Acts**: Political movements, Jewish persecution, post-war resettlement, religious appointments, and election outcomes.

— **Cluster 14 — Urbanism and Transport Infrastructure**: Demographic patterns, carnival traditions, architectural planning, pedestrian networks, and tramway integration.

To enhance interpretability and support qualitative analysis, the 10 most central articles from each cluster were identified using a medoid selection strategy[4]. Semantic distance between articles was measured using cosine similarity in the SBERT-encoded embedding space. For each cluster, the 10 articles with minimal total distance to other cluster members were selected as representative medoids.

This final set of 150 articles (15 clusters × 10 medoids) served as the semantic foundation for downstream paraphrase generation. Each medoid was manually paraphrased into 10 distinct phrasings, capturing different linguistic forms such as questions, specifications, spatial references, or temporal anchors. These paraphrases were then issued as search queries to three major search engines—Google, Bing, and DuckDuckGo. The returned results formed the empirical basis for subsequent entropy estimation, cosine similarity analysis, and cross-engine comparisons.

4.3 Query Paraphrasing Strategy

To emulate realistic user behavior during search interactions, each cluster label was transformed into ten paraphrased variants. These variants were designed to reflect different linguistic and cognitive strategies that users naturally apply—such as rewording, clarification, abbreviation, or code switching.

A local instance of LLaMA3 was used for automated paraphrase generation. The prompt enforced consistent structure and style across outputs (Listing 1).

[4] A medoid is a data point whose average dissimilarity to all other points in the cluster is minimal. Unlike centroids, medoids are actual elements of the dataset.

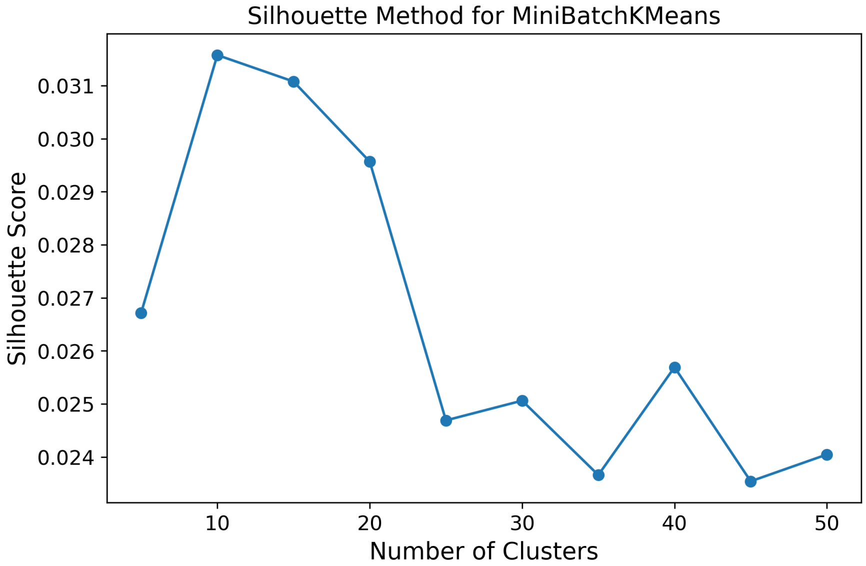

Fig. 4. Silhouette Score for different cluster counts using MiniBatch K-Means

Listing 1. Prompt format for paraphrase generation

```
Generate one paraphrase for each of these 10
user styles (each on a new line, without any
explanations or introductory phrases), for the
query "{title}":

1. Synonyms
2. Changed word order
3. Added clarification
4. Generalization
5. Question form
6. Different grammatical form
7. Abbreviation or acronym
8. Translation or code-switching (RU-EN if
possible)
9. Formal or conversational style
10. Clarification of time or place

Answer in the same language as the original
title ('title').
```

Each of the ten paraphrasing types was designed to simulate a distinct facet of human query formulation, enabling a controlled and interpretable input structure for search engine testing.

Example. For the original query *"Best Italian restaurants in New York"*, the paraphrased outputs included:

— Top-rated Italian eateries in New York

— In New York, the best Italian restaurants

— Best Italian restaurants in New York for couples

— Where to eat in New York

— What are the best Italian restaurants in New York?

— Italian dining places in New York

— NYC top Italian restos

— Where can I get good pasta in New York?

— Best Italian restaurants in NYC, 2024

4.4 Search Query Execution

Prior to large-scale querying, average response times and feasibility constraints for each search engine were benchmarked. The goal was to extract the top 50 search results—comprising *title*, *snippet*, and *URL*—for each of the 1,650 generated queries (15 clusters × 10 medoids × (10 paraphrase types + original).

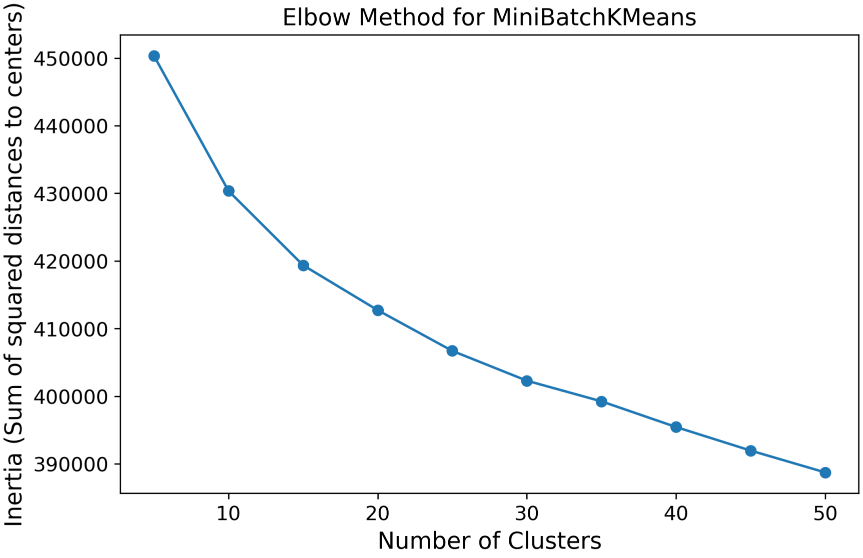

Fig. 5. Elbow method (inertia) for different cluster counts using MiniBatch K-Means

DuckDuckGo proved the most responsive and stable for scraping: using a custom headless browser pipeline, it achieved a mean latency of 21 seconds per query (including page load and DOM parsing), enabling the full dataset to be gathered in approximately 10 hours.

Bing demonstrated similar throughput and reliability. However, **Google** presented severe rate-limiting issues: after 100–200 automated queries, CAPTCHA challenges were triggered, making full-scale scraping infeasible. While paid CAPTCHA-solving APIs exist, they were rejected due to cost. Manual supervision was also considered impractical for 10+ hours of operation.

To resolve this, the **Serper API**[5] was employed to access Google results, leveraging its free tier (first 2,500 queries). The full dataset—1,650 queries—fit comfortably within this quota. However, a key limitation of Serper was that it returned only the top 10 results per query, in contrast to the 50 returned by Bing and DuckDuckGo.

Other engines such as **Yandex** and **Yahoo** were excluded due to excessive response times (estimated at over 25 hours) and low throughput on consumer hardware.

Overall, three engines were selected for comparison: DuckDuckGo, Bing, and Google. Each exhibited distinct trade-offs in terms of retrieval depth, latency, and automation tolerance. DuckDuckGo and Bing provided approximately 50 results per query, while Google's API was limited to 10. Execution time per query averaged 21 seconds for DuckDuckGo and similar values for Bing, resulting in a full scraping duration of roughly 10 hours per engine. Google scraping via headless automation was hindered by CAPTCHA interruptions every 100–200 queries. Manual CAPTCHA solving was deemed infeasible, and automated bypass solutions were cost-prohibitive. As a workaround, the Serper API was adopted to retrieve results from Google, with a daily quota sufficient for the 1,650 total queries.

The search dataset was compiled by issuing one query per paraphrase variant (11 per article) for each of the 10 medoids in all 15 clusters. Initially structured in wide format (each article with 11 paraphrased queries), the data was reshaped to long format, resulting in one row per query–engine–result triplet. The final dataset sizes were:

— **DuckDuckGo:** 77,611 rows

[5] `https://serper.dev`

Table 2. Summary statistics for each cluster after extractive summarization

Cluster	Num. Documents	Avg. Length (Words)	Avg. Length (Sentences)
0	40359	322.4	18.49
1	24431	455.4	23.10
2	38462	156.4	10.17
3	29897	423.2	22.41
4	26961	584.1	29.84
5	30721	304.3	16.52
6	26150	538.7	26.71
7	38667	238.8	13.92
8	30806	213.4	12.51
9	31957	420.4	23.37
10	33047	552.5	29.13
11	19714	529.2	27.87
12	52036	716.7	36.09
13	30862	521.5	26.41
14	45930	451.9	24.59

— **Bing:** 82,264 rows

— **Google (Serper API):** 13,983 rows

Each row includes the `title`, `snippet`, and `URL` of a retrieved result. Empty queries were automatically skipped, and DOM changes were handled through error checking and fallback logic, ensuring consistency and completeness in the collected datasets.

4.5 Entropy, Semantic Dispersion, and Relevance Score

To quantify the diversity and semantic coherence of search engine results, two complementary metrics were computed: **Shannon entropy** and **cosine distance**. These were further combined into a unified metric for comparative evaluation across engines.

1. Entropy Calculation. For each query and engine, the tokenized output (title and snippet) was aggregated across the top-n results. Shannon entropy was then computed as:

$$H(X) = -\sum_{i=1}^{k} p_i \log_2 p_i,$$

where p_i is the empirical probability of token i in the result set. Higher entropy reflects greater lexical diversity among retrieved results.

Results revealed that **DuckDuckGo** consistently yielded the highest entropy across all top-n thresholds, indicating a greater variety of lexical tokens in its search snippets. Google consistently showed the lowest entropy, with Bing falling in between (see Table 3 for $n = 50$).

2. Semantic Dispersion. To assess the semantic diversity of retrieved results, average pairwise cosine distance was computed between vectorized representations of each result. Embeddings were generated using SBERT (`all-MiniLM-L6-v2`). Each result's representation was constructed as the concatenation of its `title` and `snippet`. Formally, for N vectors $v_1, \ldots, v_N$, the average cosine distance is:

$$\bar{d} = \frac{2}{N(N-1)} \sum_{i<j} (1 - \cos(\theta_{ij})) \,,$$

where $\cos(\theta_{ij}) = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}$ is the cosine similarity between vectors i and j. This is efficiently computed via the upper triangle of the pairwise distance matrix.

3. Unified Relevance Score. To balance lexical diversity (entropy) with semantic coherence

Table 3. Summary of entropy values for top-50 results

Search Engine	Mean Entropy	Std. Dev	Max Entropy
DuckDuckGo	8.256	0.678	9.794
Bing	7.704	0.678	9.249
Google	6.215	0.765	7.478

(cosine dispersion), a unified score was computed for each query–engine pair as:

$$\text{Score} = \frac{H}{d+\varepsilon},$$

where H is the entropy, d is the average pairwise cosine distance between the embedded search result snippets and the vector representation of the original query, and $\varepsilon = 10^{-3}$ is a smoothing constant to prevent division by zero. This adjustment was particularly necessary for Google, which in some cases returned highly semantically consistent (but lexically sparse) outputs.

To enable inter-engine comparison, the final score was log-transformed:

$$\text{Final Score} = \log\left(\frac{H}{d+\varepsilon}\right).$$

This metric rewards engines that offer a balance of token variety and semantic dispersion, penalizing overly redundant or incoherent results.

Table 4 reports log-transformed scores at `top_n = 10`, the maximum consistent with Google's response depth. DuckDuckGo achieves the highest mean score across engines, while Google exhibits the largest variance due to a heavy-tailed distribution caused by sparse or inconsistent result sets.

Figure 6 shows that paraphrase robustness differs notably across engines. DuckDuckGo maintains consistently high log scores across most paraphrase types, especially for `abbreviation`, `clarification`, and `question form`, suggesting strong adaptability to surface-level reformulations. Google, despite its limited depth, displays large score fluctuations—particularly under `translation` and `synonyms`—indicating sensitivity to lexical variation. Bing exhibits lower variance overall but lacks dominance in any specific paraphrase category.

At the cluster level (Figure 7), no engine is universally superior. DuckDuckGo leads in 9 of 15 clusters, Google in 5, and Bing only in 1. Cluster 8, marked by high thematic cohesion, shows clear Google dominance, whereas Cluster 7 favors DuckDuckGo, likely due to better handling of multilingual or stylistically diverse queries.

Temporal dynamics of `top_n` (Figure 8) indicate diminishing marginal returns. DuckDuckGo peaks at $n{=}10$, followed by slight degradation and stabilization, suggesting redundancy in lower-ranked content. Bing dips at $n{=}20$, then slowly recovers—possibly reflecting over-optimization at top ranks. Google remains flat due to its hard cap at $n{=}10$, limiting its adaptive potential.

Overall, Google offers consistent but narrow results; DuckDuckGo emphasizes lexical diversity with moderate semantic noise; Bing provides a balanced but unexceptional profile. The entropy–distance score thus serves as a diagnostic of search robustness under paraphrastic stress and thematic variation.

A formal statistical evaluation was conducted to assess the effect of paraphrase type on the resulting `log(score)` values. Tests for normality (Shapiro–Wilk[6]) and homoscedasticity (Levene's test[7]) indicated non-normal distributions and heteroscedastic variance across groups, warranting the use of non-parametric methods. The Kruskal–Wallis H-test[8] returned an extremely low p-value ($p < 10^{-290}$), strongly rejecting

[6] Shapiro–Wilk test assesses whether a sample comes from a normally distributed population.

[7] Levene's test checks if multiple groups have equal variances, a key assumption in parametric tests like ANOVA.

[8] The Kruskal–Wallis test is a non-parametric alternative to one-way ANOVA, used when normality or equal variance assumptions are violated.

Table 4. Log-transformed Entropy–Distance Scores at `top_n = 10`.

Search Engine	Mean Log Score	Std. Dev.	Max Value
Bing	2.68	0.23	3.39
DuckDuckGo	2.82	0.25	3.65
Google	2.73	0.96	8.37

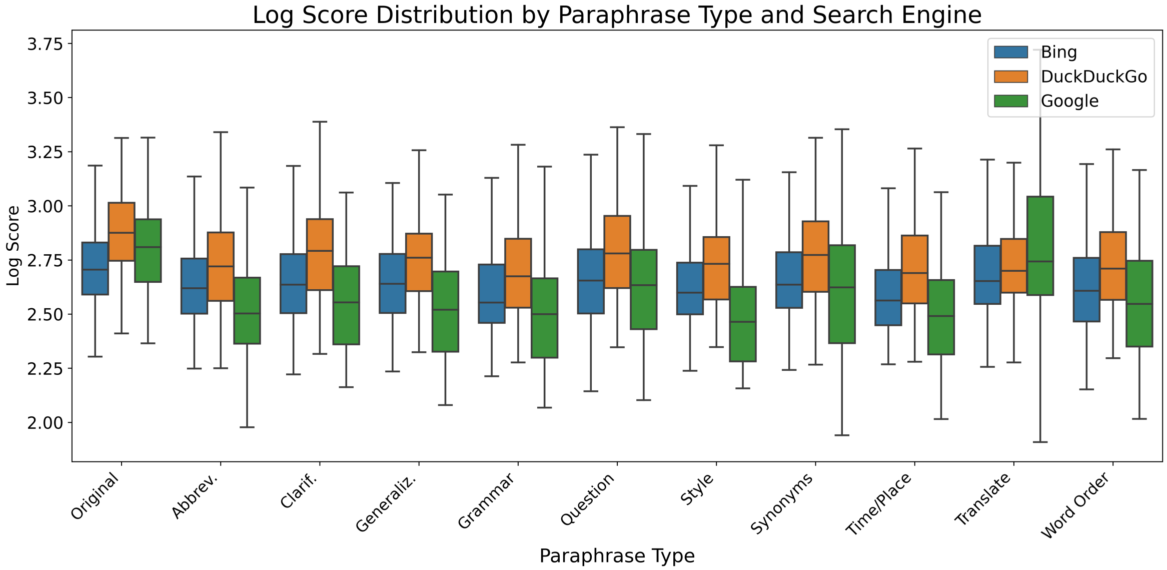

Fig. 6. Log Score Distribution by Paraphrase Type and Search Engine

the null hypothesis of equal medians across paraphrase types.

To identify specific differences, post-hoc analysis using the Tukey HSD test[9] revealed that nearly all paraphrase types significantly differ from the `original` query form (see Table 5)[10]. The most pronounced positive shifts in `log(score)` were observed for `paraphrase_translation` $(+0.13$ and above), `paraphrase_timeplace` $(+0.16$ and above), and to a lesser extent, `paraphrase_style`, `paraphrase_grammar_form`, and `paraphrase_generalization`.

These results suggest that structural paraphrases—those altering language form, tense, or scope—have substantial impact on retrieval dispersion. Conversely, paraphrases like `abbreviation`, `synonyms`, or `word_order` produce lower or even negative deviations, indicating reduced or unchanged semantic breadth and coherence.

In particular, `paraphrase_translation` yields the largest deviation from the original, implying that cross-lingual reformulation radically affects how search engines retrieve and interpret relevance. This further underscores the importance of paraphrase modeling in both multilingual search optimization and evaluation.

5 Limitations of the Study

While the study covers a wide range of experimental conditions and provides robust metrics for evaluating semantic diversity in search results,

[9] Tukey's Honestly Significant Difference (HSD) test compares all possible pairs of group means to determine which differences are statistically significant.

[10] The full set of pairwise comparisons (55 total) is available upon request

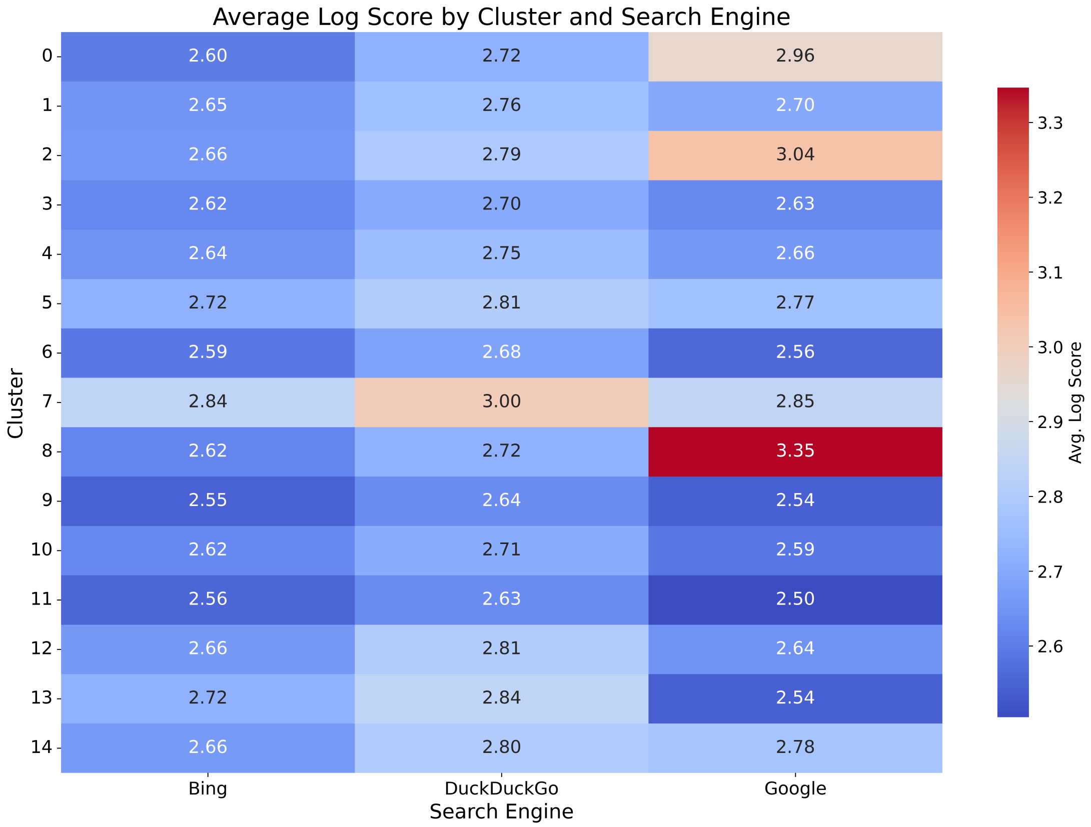

Fig. 7. Average Log Score per Cluster and Search Engine

several limitations inevitably shape its scope and interpretability.

First, the depth of result retrieval varied across search engines. Google, due to constraints imposed by the Serper API, returned only the top 10 results per query, whereas Bing and DuckDuckGo allowed retrieval of up to 50 links. This asymmetry in depth skews comparability—particularly in metrics that are sensitive to long-tail diversity, such as high `top_n` entropy evaluations.

Second, large-scale automated querying presented significant technical challenges. While Bing and DuckDuckGo could be scraped using headless browser agents, Google consistently triggered CAPTCHA protection after approximately 100–200 automated requests, effectively throttling throughput. Attempts to include other engines like Yandex and Yahoo were abandoned due to unacceptably long response times—often exceeding 25 hours per batch of paraphrased queries—rendering them impractical under current constraints.

Another important limitation concerns data sparsity. Although the raw web crawl yielded over 18 million snippets, only around 6 million were retained after filtering for meaningful content. Many discarded entries lacked coherent titles or snippets, especially among low-ranked results, which reduced the usable dataset and potentially biased the analysis toward higher-ranked, better-structured outputs.

Semantic clustering introduced its own form of heterogeneity. The 15 clusters built using medoid summaries varied in internal consistency, linguistic complexity, and topical specificity. This inconsistency led to uneven behavior across

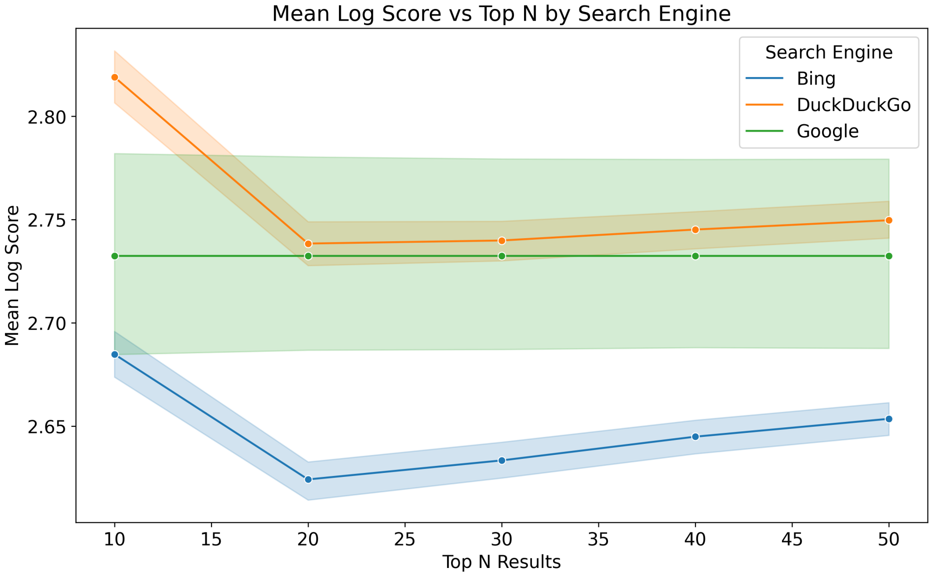

Fig. 8. Mean Log Score vs. `Top N` by Search Engine

search engines: some clusters (notably Cluster 8) produced unusually stable or unstable results depending on the engine, complicating efforts to interpret engine behavior in aggregate.

The multilingual dimension introduced further instability. Paraphrases involving code-switching or translation (particularly between Russian and English) often drifted semantically from their source prompts. These shifts, coupled with occasional encoding issues, undermined the uniformity of the paraphrase set and introduced extra variance in result quality.

A more fundamental limitation lies in the absence of human-labeled relevance judgments. The evaluation framework focused exclusively on unsupervised measures—entropy and semantic distance—rather than precision-based IR metrics grounded in user preference or utility. While this decision was intentional, it necessarily constrains the conclusions to structural properties of result distributions, rather than user-centric relevance.

Lastly, the fragility of web scraping pipelines occasionally led to result losses. Small changes in the HTML structure (DOM) of search engine results—especially in Google—sometimes broke parsing logic and silently dropped results. Although infrequent, these dropouts may have introduced slight, undetected biases favoring engines with more stable result formatting (e.g., Bing).

Despite these limitations, the study offers a replicable and data-rich framework for probing search engine behavior from a diversity-oriented perspective. Still, future work should consider augmenting entropy-based metrics with human validation, broader language coverage, and improved robustness to structural shifts in web content delivery.

6 Conclusion

This study examined how variations in paraphrase structure interact with search engine architectures to influence the diversity and quality of retrieved content. By combining entropy and cosine distance metrics into a unified evaluation framework, the research quantified the balance between lexical variety and semantic dispersion across three major search engines—DuckDuckGo, Bing, and Google. The results provide a systems-level perspective on how retrieval platforms respond to controlled linguistic reformulation, offering insights relevant

Table 5. Tukey HSD results for comparisons between the `original` query and each paraphrase type

Group 1	Group 2	Mean Diff	p-adj	95% CI	Reject
original	paraphrase_abbreviation	-0.1216	0.000	[-0.177, -0.0661]	Yes
original	paraphrase_clarification	-0.0899	0.000	[-0.1451, -0.0347]	Yes
original	paraphrase_generalization	-0.1264	0.000	[-0.1818, -0.0710]	Yes
original	paraphrase_grammar_form	-0.1355	0.000	[-0.1909, -0.0801]	Yes
original	paraphrase_question_form	-0.0872	0.000	[-0.1424, -0.0320]	Yes
original	paraphrase_style	-0.1543	0.000	[-0.2095, -0.0991]	Yes
original	paraphrase_synonyms	-0.1255	0.000	[-0.1808, -0.0702]	Yes
original	paraphrase_timeplace	-0.1644	0.000	[-0.2200, -0.1089]	Yes
original	paraphrase_translation	+0.1284	0.000	[+0.0706, +0.1862]	Yes
original	paraphrase_word_order	-0.1034	0.000	[-0.1586, -0.0482]	Yes

to both information retrieval research and the practical optimization of query formulation.

The analysis showed that DuckDuckGo consistently delivered the most stable and semantically diverse results, suggesting a stronger capacity to accommodate structural variation in user input. Bing achieved intermediate performance, while Google's outputs displayed high variance largely attributable to its fixed top-10 retrieval depth and periodic instability in HTML structure. Increasing the number of retrieved results beyond the first 10–20 yielded negligible gains in diversity, indicating that the most informative content is typically concentrated in the highest-ranked positions. This reinforces the notion that in contemporary search environments, deeper pagination contributes marginally to overall retrieval value.

Paraphrase type emerged as a critical factor shaping retrieval behavior. Structural modifications—such as translation, time or place specification, stylistic reformulation, grammatical transformation, and generalization—produced substantial changes in the entropy–distance score, reflecting shifts in semantic coverage. By contrast, surface-level variations, including synonym substitution, abbreviation, or changes in word order, generated minimal or inconsistent effects.

The weak correlation between positional rank and entropy further suggests that diversity is not inherently linked to higher ranking, challenging the assumption that top-ranked results alone capture the breadth of relevant content.

From an information retrieval standpoint, these findings underline the importance of adaptive query design and engine-specific evaluation. The entropy–distance score offers a reproducible, interpretable diagnostic for assessing retrieval robustness, enabling systematic comparison beyond relevance-based metrics. Its sensitivity to both lexical and semantic properties positions it as a practical tool for analyzing how search systems handle linguistic variation.

Moreover, the concentration of diversity within the top-20 results justifies optimization efforts at this rank range, while cross-engine differences in sensitivity highlight the need for platform-specific strategies.

Overall, the study demonstrates that controlled paraphrase variation can serve as a powerful probe for uncovering retrieval system behavior. By integrating quantitative diversity metrics with an analysis of structural query transformations, the approach advances both theoretical understanding and applied methodologies in evaluating search engine performance within dynamic and multilingual information environments.

Acknowledgments

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23486904).

References

1. **Agarwal, S., Zhang, Z., Yuan, L., Han, J., Peng, H. (2025).** The unreasonable effectiveness of entropy minimization in llm reasoning.

2. **Akhmetov, I., Gelbukh, A., Mussabayev, R. (2021).** Greedy optimization method for extractive summarization of scientific articles. IEEE Access, Vol. 9, pp. 168141–168153.

3. **Chen, M., Wang, Y., Xu, C., Le, Y., Sharma, M., Richardson, L., Wu, S.-L., Chi, E. (2021).** Values of user exploration in recommender systems. Proceedings of the 15th ACM Conference on Recommender Systems, RecSys '21, Association for Computing Machinery, New York, NY, USA, pp. 85–95.

4. **Deng, H., King, I., Lyu, M. R. (2009).** Entropy-biased models for query representation on the click graph. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, Association for Computing Machinery, New York, NY, USA, pp. 339–346.

5. **Farquhar, S., Kossen, J., Kuhn, L., Gal, Y. (2024).** Detecting hallucinations in large language models using semantic entropy. Nature, Vol. 630, No. 8017, pp. 625–630.

6. **Jiang, F., Qin, C., Yao, K., Fang, C., Zhuang, F., Zhu, H., Xiong, H. (2024).** Enhancing question answering for enterprise knowledge bases using large language models. Database Systems for Advanced Applications: 29th International Conference, DASFAA 2024, Gifu, Japan, July 2–5, 2024, Proceedings, Part IV, Springer-Verlag, Berlin, Heidelberg, pp. 273–290.

7. **Jiang, X., Liu, D., Dong, R. (2023).** Enhancing topic extraction in recommender systems with entropy regularization.

8. **Joo, E., Lee, Y.-J., Choi, H.-J. (2025).** Entropy-based sentence-level hallucination score in large language models. 2025 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 77–78.

9. **Khurana, A., Bhatnagar, V. (2022).** Investigating entropy for extractive document summarization. Expert Systems with Applications, Vol. 187, pp. 115820.

10. **Kim, G., Kim, S., Jeon, B., Park, J., Kang, J. (2023).** Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models.

11. **Kuhn, L., Gal, Y., Farquhar, S. (2023).** Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation.

12. **Lee, S. (2020).** Using entropy for similarity measures in collaborative filtering. Journal of Ambient Intelligence and Humanized Computing, Vol. 11, No. 1, pp. 363–374.

13. **Li, P., Ren, G.-J., Gentile, A. L., DeLuca, C., Tan, D., Gopisetty, S. (2023).** Long-form information retrieval for enterprise matchmaking. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, pp. 3260–3264.

14. **Markitors (2023).** SEO Statistics That Prove Its Effectiveness in 2023. `https://markitors.com/seo-statistics-that-prove-its-effectiveness-in-2023/`. Accessed: 2025-07-01.

15. **Min, S., Michael, J., Hajishirzi, H., Zettlemoyer, L. (2020).** AmbigQA: Answering ambiguous open-domain questions. **Webber, B., Cohn, T., He, Y., Liu, Y.**, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp. 5783–5797.

16. **Naqvi, S. M. R., Ghufran, M., Varnier, C., Nicod, J.-M., Javed, K., Zerhouni, N. (2024).** Unlocking maintenance insights in industrial text through semantic search. Comput. Ind., Vol. 157, No. C.

17. **Neurohive (2024).** The curse of quality saturation. `https://t.me/neurohive/1851`. Accessed: July 3, 2025.

18. **Nikitin, A., Kossen, J., Gal, Y., Marttinen, P. (2024).** Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities.

19. **OpenSource Connections (2019).** Diversity vs. Relevance in Search Systems. `https://opensourceconnections.com/blog/2019/09/05/diversity-vs-relevance/`. Accessed: 2025-07-01.

20. **Qiu, Z., Ou, Z., Wu, B., Li, J., Liu, A., King, I. (2025).** Entropy-based decoding for retrieval-augmented large language models.

21. **Search Engine Journal (2019).** Google Announces Site Diversity Change to Search Results. `https://www.searchenginejournal.com/google-site-diversity-change/311557/`. Accessed: 2025-07-01.

22. **Search Engine Journal (2023).** Google Reveals Its Methods For Measuring Search Quality. `https://www.searchenginejournal.com/google-reveals-its-methods-for-measuring-search-quality/520940/`. Accessed: 2025-07-01.

23. **Shannon, C. E. (1948).** A mathematical theory of communication. Bell System Technical Journal, Vol. 27, No. 3, pp. 379–423.

24. **Storychief (2025).** Analytics for SEO in 2025: What Metrics Matter the Most? `https://storychief.io/blog/seo-analytics-in-2025`. Accessed: 2025-07-01.

25. **Tunkelang, D. (2021).** Similarity-Sensitive Diversity. `https://dtunkelang.medium.com/similarity-sensitive-diversity-16a35d64f48c`. Accessed: 2025-07-01.

26. **Wang, Z., Duan, J., Yuan, C., Chen, Q., Chen, T., Zhang, Y., Wang, R., Shi, X., Xu, K. (2024).** Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond.

27. **Wikipedia (2025).** Entropy (information theory). `https://en.wikipedia.org/wiki/Entropy_(information_theory)`. Accessed: 2025-07-01.

28. **Writesonic (2024).** Semantic SEO Explained: How to Rank for Google's Algorithm. `https://writesonic.com/blog/semantic-seo`. Accessed: 2025-07-01.

29. **Wu, H., Zhang, Y., Ma, C., Lyu, F., He, B., Mitra, B., Liu, X. (2024).** Result diversification in search and recommendation: A survey.

30. **Xu, J., Desai, S., Durrett, G. (2020).** Understanding neural abstractive summarization models via uncertainty.

31. **Yalcin, E., Ismailoglu, F., Bilge, A. (2021).** An entropy empowered hybridized aggregation technique for group recommender systems. Expert Systems with Applications, Vol. 166, pp. 114111.

32. **Yu, T., Zhou, W., Leiyang, L., Shukla, A., Mmadugula, M., Gundecha, P., Burnett, N., Xu, A., Viseth, V., Tbar, T., Akkiraju, R., Zhang, V. (2025).** EKRAG: Benchmark RAG for enterprise knowledge question answering. **Shi, W., Yu, W., Asai, A., Jiang, M., Durrett, G., Hajishirzi, H., Zettlemoyer, L.**, editors, Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing, Association for Computational Linguistics, Albuquerque, New Mexico, USA, pp. 152–159.

33. **Zhou, C., You, W., Li, J., Ye, J., Chen, K., Zhang, M. (2023).** INFORM : Information entropy based multi-step reasoning FOR large language models. **Bouamor, H., Pino, J., Bali, K.**, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, pp. 3565–3576.

34. **Zhou, D. X., Liu, L., Anubhai, A., Shandilya, M., Sigalas, S., Wang, W. Y., Huang, Z. (2023).** Beyond accurate answers: Evaluating open-domain question answering in enterprise search. Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR '23, Association for Computing Machinery, New York, NY, USA, pp. 308–312.

35. **Zuccon, G., Azzopardi, L. (2016).** Using the quantum probability ranking principle to rank interdependent documents. Proceedings of the 38th European Conference on Information Retrieval, pp. 357–369. Accessed: 2025-07-01.

Article received on 01/06/2025; accepted on 16/10/2025.
**Corresponding author is Alexander Gelbukh.*