A Comprehensive Survey on Interpretable Sentence-Level Relevance and Importance Modeling: From Statistical Heuristics to Transformer-Based and Correlation-Guided Frameworks

Naresh Kumar Yalla[1], V. Kamakshi Prasad[2]
[1,2]*Department of Computer Science and Engineering Jawaharlal Nehru Technological University, Hyderabad, India*

***Abstract*—Sentence-level relevance scoring underpins modern information retrieval, extractive summarization, question answering, and decision-support systems, where ranking quality directly influences interpretability, reliability, and downstream reasoning. Research in this area has evolved from statistical heuristics and graph-based centrality to neural scoring–selection architectures, transformer encoders, and emerging large language model (LLM)-assisted estimation. Despite substantial gains in semantic expressiveness, persistent structural challenges remain, including fragmented operational definitions of relevance, redundancy amplification arising from correlated signals, limited transparency of factor interactions, evaluation protocols overly dependent on lexical overlap, and instability under domain shift or prompt variation.**

This survey provides a structured and governance-oriented synthesis of sentence-level relevance modeling. We formalize relevance as a context-conditioned, multi-factor scoring function and analyze how lexical, semantic, structural, discourse, and informativeness signals are integrated across paradigms. Beyond chronological review, we organize existing methods through the lenses of multi-factor interaction modeling, correlation-aware signal governance, and reliability-sensitive evaluation. Particular emphasis is placed on redundancy control, feature decorrelation, representation stability, and faithfulness-aligned validation mechanisms essential for preventing proxy dominance and unstable attribution.

By comparatively analyzing modeling assumptions and evaluation practices across tasks, this survey identifies systemic gaps and articulates research directions toward explainable, correlation-governed, and stability-calibrated sentence-level relevance frameworks. We argue that advancing the field requires repositioning relevance modeling from performance-centric optimization toward governance-aware and reliability-sensitive design suitable for real-world deployment.

***Index Terms*—Sentence-level relevance, extractive summarization, information retrieval, explainable AI, correlation-aware modeling, faithfulness, transformer models, large language models**

I. INTRODUCTION

A. Sentence-level relevance estimation is a fundamental modeling primitive in modern information systems. It underlies doc-amend ranking, extractive and hybrid summarization, question answering, conversational agents, legal document analytics, and large language model (LLM)-driven reasoning pipelines. In both academic research and industrial deployment, relevance scoring determines which information becomes visible, trusted, and actionable. Consequently, instability or misalignment in sentence-level scoring propagates to higher-level tasks, affecting decision reliability, fairness, and factual robustness. Early approaches operationalized relevance through statistical and heuristic signals such as term frequency, sentence position, cue phrases, and handcrafted feature combinations [1], [2]. These models offered interpretability and computational efficiency but relied heavily on lexical overlap and manually tuned weights. Graph-based centrality frameworks subsequently reframed salience as a relational property emerging from sentence similarity networks, where eigenvector propagation captured mutual reinforcement among sentences [3], [4]. This marked an important conceptual shift:

relevance became dependent on inter-sentence interactions rather than isolated sentence attributes. The neural paradigm further transformed sentence scoring into a learned contextual representation problem. Joint scoring–selection models demonstrated that importance is dynamic and conditioned on previously selected content, highlighting the contextual and sequential nature of relevance [5]. Sentence pair scoring frameworks unified diverse tasks including answer selection, entailment, and dialogue ranking under a common relevance function f2(s0, s1), emphasizing crosstask transferability and statistically grounded evaluation [6].

Transformer-based architectures extended these capabilities by enabling scalable bidirectional contextual encoding tailored for sentence scoring [7]. More recently, LLM-driven reranking and prompting strategies have further expanded relevance estimation into generative reasoning systems.

Despite these advances, several structural challenges remain unresolved. First, conceptual ambiguity persists: relevance is often conflated with salience, importance, or utility without a unified operational definition [8]. Second, modern systems integrate lexical, semantic, structural, and discourse level signals without explicit governance of feature interaction or redundancy, potentially leading to multicollinearity and unstable explanations. Emerging explainable AI approaches attempt to address this limitation through correlation-aware feature weighting and group-level attribution mechanisms [9], [10], yet systematic integration into sentence-level relevance modeling remains limited. Third, evaluation frameworks frequently rely on lexical-overlap metrics such as ROUGE, which correlate weakly with factual faithfulness and reliability. Empirical analyses demonstrate that neural summarization systems often generate hallucinated or unsupported content, even when achieving high automatic scores [11].

These limitations motivate a structured re-examination of sentence-level relevance modeling beyond performance-centric evaluation. There is increasing need for multi-factor integration frameworks, correlation-aware redundancy control mechanisms, and evaluation paradigms grounded in explainability, faithfulness, and robustness. By synthesizing developments across statistical, graph-based, neural, and LLM-driven paradigms, this survey advances a comprehensive perspective on sentence-level relevance as an explainable, correlation governed, and reliability-aware modeling problem.

II. CONCEPTUAL FOUNDATIONS OF SENTENCE-LEVEL RELEVANCE

Sentence-level relevance constitutes a foundational construct across information retrieval, extractive summarization, ranking systems, and representation learning frameworks [8], [12]. Although the terms relevance, salience, importance, and utility are often used interchangeably in applied literature, they represent distinct theoretical abstractions with different operational implications. Clarifying these distinctions is essential for building stable, interpretable, and multi-factor sentence scoring models.

In its most general abstraction, sentence-level relevance may be represented as a scoring function:

$$f: S \times C \rightarrow R, \quad (1)$$

where S denotes the sentence space, C denotes contextual or task-specific representations (e.g., queries, document themes, prior selections, or objective constraints), and R denotes the set of real numbers representing the real-valued scoring domain. The output of f is a scalar relevance estimate whose magnitude reflects the degree of alignment between a sentence and a defined objective. While the internal structure off varies across paradigms ranging from statistical aggregation and graph propagation to neural encoding and marginal-gain optimization the abstraction of relevance as relational alignment remains invariant.

A. Relevance

Relevance is traditionally defined as the degree to which a sentence satisfies an external information need or task objective. Early extractive summarization systems approximated relevance using surface-level statistical indicators such as term frequency, positional heuristics, and cue phrases [1], [2], [13]. Feature aggregation and fuzzy-based methods extended these mechanisms while incorporating redundancy mitigation strategies [4], [14].

Graph-based centrality approaches reframed relevance as relational prominence within sentence similarity networks [3], [15]. Through eigenvector-style reinforcement, sentences accrue importance via structural connectivity, emphasizing mutual similarity.

Sentence-pair scoring frameworks generalized relevance as a learned relational function f (s0, s1), applicable across entailment, ranking, and question answering tasks [6]. Joint scoring–selection models conditioned sentence importance on previously selected content using marginal-gain optimization [5].

Transformer-based architectures enhanced contextual interaction modeling through attention mechanisms, enabling dynamic alignment between sentence representations and task objectives [7], [16]. Across these paradigms, relevance consistently represents externally conditioned relational alignment.

B. Salience

Salience refers to the intrinsic structural prominence of a sentence within a document. Unlike relevance, which is externally conditioned, salience emphasizes internal centrality or informational prominence. Centrality-based methods equate salience with graph-theoretic importance derived from similarity networks [3]. Neural extractive frameworks reinterpret salience as a latent representation optimized under supervision [5]. Domain-specific adaptations further refine salience modeling in structured corpora such as legal documents [17].

C. Importance

Importance provides a broader abstraction integrating multiple interacting informational factors. Multi-factor explainable frameworks formalize importance as a structured combination of relevance, novelty, and informativeness [18], [19]. Broader explainable AI literature reinforces the necessity of factor-level interpretability in scoring mechanisms [10].

D. Utility

Utility generalizes importance toward task-optimized and user-sensitive objectives. Explainable selection frameworks incorporate controllable thresholds, interaction modeling, and governance constraints to regulate sentence extraction [18]. Faithfulness-sensitive evaluation research further highlights the need to balance alignment with factual consistency [11], [20].

E. Correlation-Aware Extensions

Recent research introduces correlation-aware feature governance mechanisms to mitigate redundancy, feature collapse, and representational instability [9], [21]. These directions emphasize structural stability alongside alignment and coverage optimization, marking a transition toward governed and explainable sentence scoring paradigms.

F. Conceptual Distinction Summary

For theoretical clarity:

- Relevance: externally conditioned relational alignment under an objective.
- Salience: intrinsic structural prominence within a document.
- Importance: multi-factor integration of relevance, novelty, and informativeness.
- Utility: task-optimized operational value under contextual constraints.

These distinctions establish a principled theoretical foundation for explainable, multi-factor, and correlation-aware sentence-level relevance modeling.

III. EVOLUTION OF SENTENCE-LEVEL SCORING METHODS

Sentence-level scoring has undergone a substantial methodological transformation over the past two decades. What began as heuristic feature aggregation evolved into graph-based relational modeling, followed by neural representation learning and transformer-based conditional scoring. More recently, large language models (LLMs) have introduced reasoning driven and provenance-aware relevance estimation. Beyond performance improvements, each transition reflects a deeper conceptual shift redefining the objective of relevance, the representation of sentences, the supervision signal, and the degree of interpretability embedded within the scoring process [8], [12].

A. Statistical and Feature-Based Approaches

Early sentence scoring systems operationalized relevance as a linear combination of handcrafted surface features:

$$\text{Score}(s_i) = \sum_{k=1}^{K} w_k f_k(s_i), \quad (2)$$

where f_k denotes manually engineered features and w_k represents heuristic or fixed weights.

Comparative analyses identified dominant feature families, including term frequency, TF–IDF, positional heuristics, title similarity, cue phrases, proper noun density, and sentence length [1], [2], [13]. Feature bundles generally outperformed individual

heuristics but remained corpus-dependent and sensitive to positional bias [2]. Extensions incorporated fuzzy-logic aggregation, multi-document heuristics, and redundancy-aware selection mechanisms [4], [14].

During this era, supervision was minimal or indirect, and evaluation relied heavily on lexical overlap metrics such as ROUGE. Relevance was thus approximated through observable lexical proxies rather than learned semantic alignment.

B. Graph-Based Centrality Models

Graph-based methods reconceptualized relevance as a relational property. Under this view, a sentence is important if it is similar to many other important sentences.

Lex Rank formalized this framework by constructing a sentence similarity graph and computing eigenvector-based centrality scores using PageRank-style propagation [3]. GETS style pipelines extended this principle through lexical weighting and redundancy-aware refinement [4], [15]. Comparative evaluations confirmed that centrality methods remain competitive baselines, though highly sensitive to similarity design and thresholding strategies [2].

The conceptual shift here is decisive: importance becomes a network-propagated structural property rather than an independently computed scalar score. However, optimization objectives remained indirectly tied to lexical similarity signals, and redundancy amplification persisted when near-duplicates existed [3].

C. Neural and Transformer-Based Models

Neural approaches replaced handcrafted heuristics with learned representations and task-aligned optimization objectives, enabling semantically grounded relevance estimation.

1) Joint Scoring and Selection (State-Aware Relevance): Joint scoring–selection frameworks modeled relevance as marginal utility conditioned on previously selected content [5]. By optimizing step-wise gain functions relative to a partial summary state, these models reframed relevance as history dependent rather than intrinsic, integrating redundancy control directly into the objective function.
2) Explainable and Controllable Neural Selection: Explainability emerged through decomposable selection architectures. ESCA-style frameworks explicitly model interactions among relevance, novelty, and informativeness while exposing controllable inference parameters [18]. Broader explainable modeling frameworks further emphasize structured and interpretable relevance estimation [10], [19].

This period marks a tension between performance gains and interpretability: while neural models improved semantic representation, they often reduced transparency compared to feature-based methods.

3) Transformer-Based Sentence Representations: Transformer architectures strengthened semantic alignment through contextual encoding and bidirectional interaction modeling. Transcormer explicitly treats sentence scoring as a primary modeling objective [7], while unified sentence-pair scoring paradigms generalize relevance as a transferable function across tasks [6]. Global encoding techniques further improve contextual representation quality [16].

A critical methodological shift during this era concerns supervision signals. Unlike earlier heuristic methods, neural models optimized differentiable objectives often proxy metrics such as ROUGE raising concerns about alignment with factual correctness and semantic faithfulness.

D. Faithfulness and Reliability Extensions

Faithfulness-sensitive research highlighted limitations of lexical overlap metrics for capturing semantic validity [11]. Long-premise reasoning frameworks further emphasized the need for consistency-aware relevance estimation in extended contexts [20]. These developments mark a shift from performance-centric optimization toward reliability aware evaluation.

E. Correlation-Aware and Governance-Driven Extensions

Recent research introduces correlation-aware feature governance mechanisms to mitigate redundancy amplification and feature collapse. Dual correlation reduction frameworks promote structural stability in representation learning [21], while adaptive feature-weighting mechanisms incorporating Pearson correlation enhance explainability and robustness [9].

This transition reflects an emerging paradigm: beyond maximizing alignment or coverage, sentence scoring increasingly integrates structural stability, redundancy

control, interpretability, and governance constraints.

F. Large Language Model-Based Relevance Estimation

Large language models extend sentence scoring beyond static encoders by leveraging instruction-following, reasoning, and contextual attribution.

LLMs enable prompt-based graded relevance scoring and rapid domain adaptation, though such approaches remain sensitive to prompt design and calibration. In high-stakes domains such as legal summarization, structured clustering and representation learning emphasize traceable and context-aware relevance modeling [17].

Collectively, the evolution from feature heuristics to Lambasted reasoning reflects a broader conceptual shift: sentence level relevance is no longer treated as a static lexical similarity score but as a dynamic, semantically grounded, multi-factor, and increasingly governance-aware construct.

IV. MULTI-FACTOR RELEVANCE MODELING

Sentence-level relevance is inherently multi-dimensional. Across summarization, retrieval, ranking, and entailment-oriented tasks, no single signal lexical overlap, semantic similarity, structural prominence, or informational density adequately captures the full notion of relevance [8], [12]. Contemporary systems therefore integrate heterogeneous evidence sources into composite scoring mechanisms. The shift toward multi-factor modeling reflects a broader recognition: relevance is a structured construct emerging from interacting and potentially correlated signals rather than a single similarity score.

A. Types of Relevance Signals

1) Lexical Signals: Lexical signals include term frequency, TF–IDF weights, keyword matching, title similarity, namedentity density, and n-gram overlap [1], [2], [13]. Such features remain strong baselines in extractive summarization and sentence ranking [4], [6]. However, reliance on surface overlap limits robustness under paraphrasing and semantic reformulation.
2) Semantic Signals: Semantic signals capture contextual meaning and relational alignment using learned representations. Transformer-based encoders and cross-encoder architectures model fine-grained token interactions to improve semantic grounding [7], [16]. Joint scoring–selection frameworks further demonstrate that semantic contribution is context dependent [5]. Faithfulness-oriented modeling reinforces the need for semantic validity beyond lexical coincidence [11], [20].
3) Structural Signals: Structural signals encode document organization, including sentence position and section boundaries [2]. Graph-based centrality approaches implicitly capture structural prominence through connectivity patterns [3]. Domain-adaptive systems further exploit structural cues in specialized corpora such as legal documents [17].
4) Discourse and Interaction Signals: Discourse-aware modeling captures redundancy, novelty, coverage, and marginal gain relative to selected content. State-aware extraction frameworks explicitly integrate redundancy suppression into relevance estimation [5]. Explainable selection models decompose interaction factors into relevance, novelty, and informativeness [18], [19].
5) Informativeness Signals: Informativeness measures information density and contribution independent of redundancy. Multi-factor explainable frameworks explicitly model informativeness as a distinct component influencing sentence selection [18]. This distinction ensures that sentences are evaluated not only for alignment but also for substantive content contribution.
6) Uncertainty and Reliability Signals: Emerging research highlights the importance of confidence estimation and reliability-aware modeling. Faithfulness evaluation and correlation-aware stability analysis underscore the need to assess not only relevance magnitude but also prediction robustness [11], [21].

B. Feature Fusion and Objective-Level Integration

1) Linear Aggregation: Traditional fusion relies on weighted summation: $\text{Score}(s_i) = \Sigma\, w_k f_k(s_i)$, (3)k

where weights may be manually tuned or heuristically assigned [2], [14]. Although interpretable, this approach ignores inter-factor correlation.

2) Graph-Based Propagation: Graph-based fusion integrates signals via relational consensus rather

than explicit summation [3]. Extended frameworks incorporate lexical weighting and clustering for scalability [4], [15].

3) Neural Fusion and Multi-Objective Optimization: Neural architectures implicitly fuse signals within learned representations [5], [7]. Importantly, modern systems optimize multi-objective functions such as marginal gain, coverage, and redundancy suppression rather than isolated similarity scores. Unified scoring paradigms emphasize cross-task transferability of learned relevance functions [6].
4) Correlation-Aware Fusion: Aggregating correlated signals without governance can inflate importance attribution. Dual-level correlation reduction stabilizes representations [21], while adaptive feature-weight explanation mechanisms identify non-redundant factor groups prior to weighting [9]. These methods extend multi-factor modeling toward structured governance.

C. Dynamic and Adaptive Weighting

Static weights assume constant factor importance. Contemporary systems adopt adaptive strategies:

- Data-Driven Learning: Factor weights emerge through task-aligned objectives such as marginal ROUGE gain [5].
- History-Aware Adjustment: Relevance is conditioned on prior selections [5].
- Controllable Thresholding: Explainable architectures expose inference-time control parameters [18].
- Correlation-Governed Adaptation: Factor weights are adjusted based on inter-factor dependencies and stability considerations [9], [21].

D. Evaluation Alignment

Multi-factor relevance modeling must align with evaluation beyond lexical overlap. Faithfulness-sensitive metrics and semantic consistency checks provide complementary validation signals [11]. Stability analysis and reproducibility reporting further strengthen governance-aware scoring.

Overall, multi-factor relevance modeling represents a transition from isolated heuristics to structured, interaction-aware, reliability-sensitive, and correlation-governed architectures.

V. CORRELATION-AWARE RELEVANCE MODELING

Sentence-level relevance does not emerge independently across sentences or scoring factors. Instead, it is shaped by multiple forms of dependence:

- Redundancy: repeated informational content across sentences,
- Correlation: statistical dependence among features or representations,
- Dependency: logical or semantic influence (e.g., entailment).

While these constructs are related, they require distinct modeling strategies. Correlation-aware relevance modeling governs inter-sentence redundancy, inter-feature multicollinearity, inter-factor interactions, and representation-level instability to prevent redundancy amplification and unstable importance attribution.

A. Inter-Sentence Correlation and Redundancy

Graph-based centrality models encode inter-sentence similarity via adjacency matrices and eigenvector propagation [3]. However, dense clusters of near-duplicate sentences may amplify correlated signals and dominate importance rankings.

Sequential extractive frameworks mitigate this issue through marginal gain modeling. NEUSUM defines incremental contribution as:

$$g(S_i \mid S_{t-1}) = r(S_{t-1} \cup \{S_i\}) - r(S_{t-1}), \quad (4)$$

where sentence value is conditioned on previously selected content [5]. This explicitly models redundancy as dependency relative to summary state.

Explainable selection architectures further encode pairwise sentence influence through structured interaction matrices [18]. These approaches shift from independent scoring to dependency-aware modeling.

Limitation. Most redundancy detection remains similarity driven and may underperform under semantic paraphrase or entailment-level overlap [3].

B. Inter-Factor Correlation in Multi-Factor Relevance

Multi-factor relevance modeling integrates lexical, semantic, structural, novelty, and informativeness signals [12], [18]. However, these factors are not statistically independent. For example:

- Positional bias may correlate with lexical density,
- Novelty may inversely correlate with relevance,
- Informativeness may correlate with entity density.

Aggregating correlated factors without governance inflates proxy indicators and destabilizes explanation

outputs. Adaptive Feature Weight Genetic Explanation (AFWGE) addresses this by identifying non-redundant feature groups through Pearson correlation analysis before weight assignment [9].
Thus, correlation-aware modeling must explicitly regulate inter-factor dependencies to preserve interpretability and prevent factor dominance.

C. Relational Propagation and Representation Stability

Representation learning research demonstrates that correlated samples and features may induce representation collapse. Dual correlation reduction introduces complementary mechanisms at both sample and feature levels to mitigate this instability [21].
Applied to sentence relevance modeling, this implies:

- Sentence embeddings may converge toward degenerate consensus under redundant supervision,
- Graph propagation may amplify correlated noise,
- Orthogonalized latent spaces improve discriminability and robustness.

Correlation regularization therefore stabilizes relational learning and prevents consensus amplification artifacts.

D. Optimization-Level Governance

Correlation control can be incorporated at the optimization level through:

- Covariance penalties among factor representations,
- Orthogonality constraints across latent dimensions,
- Multi-objective trade-offs balancing relevance, novelty, and informativeness.

State-aware scoring frameworks implicitly perform dependency-sensitive optimization [5], while explainable architectures expose factor interactions prior to aggregation [18]. However, explicit covariance-aware objectives remain underexplored in sentence-level relevance modeling

E. Stability, Domain Robustness, and Reproducibility

Correlation patterns vary across domains and datasets. Positional signals may dominate in news corpora but weaken in legal or biomedical contexts [17]. Unified sentence scoring frameworks emphasize multi-run evaluation, confidence interval reporting, and development-test correlation analysis to mitigate stochastic instability and dataset bias [12].
Correlation-aware modeling therefore extends beyond redundancy suppression toward reproducibility governance.

F. Metric Alignment and Reliability Governance

Faithfulness research demonstrates weak correlation between lexical overlap metrics and factual correctness [11]. Long-premise reasoning further highlights semantic consistency requirements in extended contexts [20].
This reveals a higher-order governance requirement: the correlation between scoring signals and evaluation metrics must itself be examined. Relevance signals should align with semantically grounded evaluation criteria rather than surface proxies.
In summary, correlation-aware relevance modeling represents a structural evolution from similarity-based redundancy filtering toward principled multi-level governance. By integrating inter-sentence redundancy control, inter-factor decorrelation, representation stability, optimization-level regularization, and metric alignment, modern systems move toward stable, interpretable, and reliability-sensitive sentence-level relevance estimation.

VI. EXPLAINABILITY, FAIRNESS, AND RELIABILITY IN RELEVANCE MODELS

As sentence-level relevance modeling transitions from heuristic scoring to neural and large language model (LLM)based architectures, interpretability, fairness, and robustness have emerged as central research concerns. High predictive accuracy does not guarantee faithful reasoning, unbiased ranking, or stability under perturbation. Modern relevance systems must therefore be evaluated not only for task performance but also for transparency, correlation governance, and reliability [8], [10], [12].

A. Intrinsic and Post-hoc Explainability

Explainability in relevance modeling can be broadly categorized into intrinsic (architectural) and post-hoc (external) mechanisms [10].
Intrinsic Explainability. Intrinsic approaches embed interpretability directly within the scoring or selection architecture. Interaction-based selection frameworks explicitly decompose sentence contribution into relevance, novelty, and informativeness via structured interaction matrices and controllable thresholds [18], [19]. Sequential marginal-gain models such as

NEUSUM operationalize conditional contribution relative to previously selected content, exposing procedural transparency through state-aware relevance estimation [5].

Intrinsic explainability typically provides:

a. Architectural transparency and explicit factor decomposition,
b. Inter-sentence interaction visibility (novelty/coverage effects),
c. Controllable decision parameters for governance [18].

Post-hoc Explainability. Post-hoc methods attempt to explain black-box scoring decisions without changing model structure. Counterfactual and attribution-based methods estimate sensitivity of predictions to feature perturbations and can identify correlated proxy dependence [9]. This is especially relevant in multi-factor relevance scoring, where multicollinearity may inflate attribution and produce unstable explanations. Representation-level dominance of correlated factors further motivates decorrelation-based stabilization [21].

Two recurring risks arise:

d. Plausible but unfaithful explanations: the explanation reads well but does not reflect the true computation.
e. Proxy-driven explanations: highlighted factors are correlated surrogates rather than causal drivers [9].

B. Explanation Evaluation: Faithfulness, Stability, and Correlation Governance

A key distinction in trustworthy relevance modeling is between:

a. Faithfulness: whether an explanation reflects the model's internal decision logic,
b. Plausibility: whether an explanation appears reasonable to human observers.

Empirical evidence from summarization shows that strong overlap-based performance does not imply factual correctness or faithful content selection [11]. Long-premise reasoning work further highlights degradation under extended contexts [20], implying that explanation validity must be evaluated under realistic long-document conditions.

Beyond qualitative inspection, explanation evaluation in relevance models should include:

c. Perturbation faithfulness tests: whether removing or perturbing the claimed-important features changes scores materially (counterfactual sensitivity) [9].
d. Stability tests: whether attributions remain consistent across runs, seeds, or minor input reformulations [12].
e. Correlation governance checks: whether explanations collapse onto highly correlated proxy groups rather than distinct causal factors [9], [21].

This evaluation layer is critical because correlation structures can cause explanation instability even when predictive performance remains high [12].

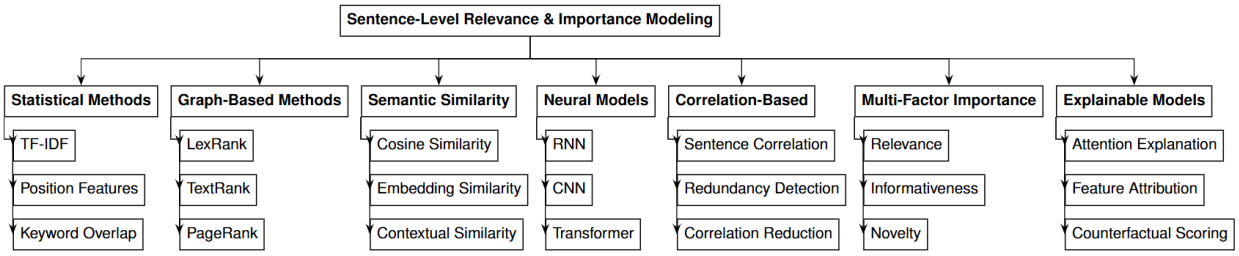

Fig. 1. Taxonomy of Sentence-Level Scoring and Relevance Modeling.

C. Bias and Fairness in Sentence-Level Ranking

Sentence-level ranking systems inherit biases from training data distributions, feature design, and architectural inductive biases [10].

Position Bias. Structured datasets exhibit lead bias, where early sentences are disproportionately labeled important. Graph-based and neural systems may reduce but do not eliminate this artifact [3], [5]. Over-

reliance on positional signals can distort ranking behavior in domains lacking rigid structural conventions.

Proxy Feature Bias and Correlated Dominance. Highly correlated features (lexical overlap, entity density, topical frequency, positional indicators) can function as proxies for importance. Without explicit governance, proxy dominance inflates attribution and reduces interpretability reliability [9]. Representation-level correlation dominance can also reduce discriminability, motivating decorrelation strategies [21].

Dataset-Induced Bias and Generalization Risk. Unified scoring research emphasizes multi-run evaluation and confidence interval reporting, highlighting that single-run ranking claims may be statistically unstable [12]. Domain-specific settings (e.g., legal summarization) further amplify fairness risks, including systematic under-ranking of nuanced arguments and over-selection of dominant narrative structures [17].

D. Reliability and Robustness in Neural and Transformer Scoring

Transformer scoring and sentence-pair scoring frameworks improve semantic alignment but often reduce transparency. Transcormer explicitly treats sentence scoring as a primary objective [7], while sentence-pair scoring abstracts relevance as a transferable function across tasks [6]. These paradigms increase the need for reliability governance under:

a. cross-domain transfer,
b. distribution shift,
c. correlated feature dominance,
d. long-context evaluation [12], [20].

E. Hallucination and Robustness in LLM-Based Relevance Estimation

LLMs are increasingly employed for sentence-level scoring via prompting, reranking, or rationale generation. While powerful, they introduce reliability risks beyond conventional neural scoring.

Unsupported Rationales and Faithfulness Risk. LLMs may assign high relevance to unsupported claims or produce plausible but incorrect rationales. Faithfulness analyses show that overlap metrics fail to detect such distortions [11]. Without entailment or consistency-sensitive validation, LLM-based scoring may promote semantically inconsistent content.

Prompt Sensitivity and Ranking Instability. LLM judgments can be sensitive to prompt phrasing and formatting, creating reproducibility and stability concerns. Stability-sensitive evaluation and multi-run reporting are therefore required [12]. Long-Context Degradation. Sentence-level relevance estimation for long documents must be validated under long-premise and multi-hop reasoning conditions [20].

F. Governance Checklist for Trustworthy Relevance Scoring

To operationalize explainability, fairness, and reliability in sentence-level relevance modeling, governance should incorporate:

a. Statistical stability: multi-run evaluation and confidence intervals [12],
b. Explanation faithfulness: counterfactual/perturbation testing for claimed-important factors [9],
c. Correlation governance: detect and mitigate correlated proxy dominance at feature and representation levels [9], [21],
d. Faithfulness-sensitive validation: semantic consistency checks beyond lexical overlap [11], [20],
e. Interpretability controls: factor-level decomposition and controllable thresholds where feasible [18], [19],
f. Bias auditing: positional/proxy bias tests and domain transfer stress testing [3], [17].

In summary, explainability, fairness, and reliability are structural requirements in modern relevance modeling. As models become more expressive and opaquer, correlation aware governance, faithfulness-aligned validation, and stability-sensitive evaluation become essential for trustworthy sentence-level relevance estimation [10], [12].

VII. COMPARATIVE SYNTHESIS, STRUCTURAL.LIMITATIONS, AND OPEN CHALLENGES

This section synthesizes methodological evidence across feature-based, graph-based, neural/transformer, and LLMdriven sentence scoring paradigms. Rather than treating these families as isolated approaches, we

analyze their structural assumptions, optimization strategies, interpretability mechanisms, evaluation practices, and governance maturity.

Although semantic expressiveness has improved substantially over the past two decades, recurring structural weaknesses persist particularly in unified relevance formalization, explicit factor interaction modeling, correlation governance, faithfulness validation, and statistical stability reporting.

A. Comparative Structural Synthesis

Feature-Based Models. Feature-based systems provide interpretability and computational efficiency by modeling explicit signals such as TF–IDF, position, cue phrases, and entity density [1]. Their transparency facilitates debugging and domain adaptation. However, they assume additive independence among factors and rely on heuristic weighting, limiting nonlinear interaction modeling and semantic robustness.

Graph-Based Centrality Models. Graph-based approaches redefine relevance as relational prominence emerging from similarity networks [3]. Eigenvector-style propagation partially mitigates redundancy and provides structural interpretability.

Nevertheless, similarity bottlenecks persist: lexical similarity inadequately captures semantic equivalence and discourse dependency [4]. Centrality may amplify correlated clusters, and threshold sensitivity affects stability.

Neural and Transformer-Based Models. Neural architectures substantially improve contextual modeling and semantic alignment. Joint scoring–selection frameworks introduce marginal gain optimization [5], while sentence-pair scoring generalizes relevance across tasks [6]. Transformer-based scoring models treat relevance as a primary modeling objective [7]. However, factor interactions are implicitly learned rather than explicitly governed, reducing interpretability and embedding metric bias when trained using ROUGE-based supervision [5]. Stability reporting remains inconsistent [12].

LLM-Based Relevance Estimation. LLMs extend sentence scoring via instruction-following and contextual reasoning. They enable zero-shot ranking and rubric-guided evaluation but introduce prompt sensitivity, hallucination risk, and nondeterminism [11]. Standardized stability and evaluation protocols remain underdeveloped [12].

B. Comparative Evaluation Across Paradigms

Table I summarizes structural characteristics across major paradigms.

The table highlights a consistent pattern: modeling expressiveness has increased, but interpretability, correlation governance, and standardized stability evaluation have not progressed proportionally

C. Taxonomy of Structural Limitations

The evolution of sentence-level relevance modeling can be taxonomized across five dimensions:

1) Representation Level: lexical → relational → contextual → generative.
2) Interaction Modeling: additive weighting → graph propagation → marginal utility → instruction conditioned reasoning.
3) Governance Mechanisms: none → redundancy filtering → implicit regularization → partial decorrelation.
4) Evaluation Paradigm: overlap-based → marginal gain → semantic entailment → stability-aware evaluation.
5) Interpretability Level: explicit features → structural centrality → latent attention → opaque prompting.

This taxonomy reveals that governance sophistication lags behind modeling sophistication, particularly in correlation regulation and statistical rigor.

D. Persistent Structural Limitations Across paradigms, several systemic limitations remain:

1. Fragmented Operational Definitions of Relevance. Relevance is alternately defined as centrality [3], marginal gain [5], lexical similarity [1], or implicit neural utility. A unified, task-agnostic multi-factor operational definition remains underdeveloped.
2. Limited Explicit Modeling of Multi-Factor Interactions. Most systems assume additive independence or implicit nonlinear fusion. While interaction-aware architectures exist [18], systematic integration into general relevance modeling is sparse.

Table I. Comparative Structural Characteristics of Sentence-Level Relevance Paradigms

Paradigm	Interpretability	Semantic Robustness	Correlation Governance	Faithfulness Alignment	Stability Reporting	Computational Cost
Feature-Based	High	Low	Weak	Weak	Rare	Low

Graph-Based	Moderate	Low–Moderate	Implicit	Weak	Rare	Low–Moderate
Neural/Transformer	Low	High	Implicit	Limited	Inconsistent [12]	High
LLM-Based	Very Low	Very High	Weak	Unverified [11]	Limited	Very High

3. Weak Correlation Governance. Redundancy is often handled during selection rather than signal construction [5]. Correlated features may inflate attribution unless explicitly regulated [9]. Representation-level collapse under correlated supervision further motivates decorrelation [21].
4. Evaluation Misalignment with Faithfulness. ROUGEbased improvements do not guarantee semantic correctness [11]. Entailment-sensitive evaluation remains inconsistently adopted [20].
5. Insufficient Stability and Statistical Reporting. Multirun variance, seed sensitivity, and weak development–test correlation is inconsistently disclosed [12], limiting reproducibility.
6. Limited Domain-Constrained Governance. Highstakes contexts such as legal summarization require traceable and entailment-consistent scoring [17]. Domain-specific governance mechanisms remain under-integrated.
7. Emerging Reliability Challenges in LLM-Based Scoring. Prompt sensitivity, stochastic decoding, and hallucinated rationales introduce instability dimensions that lack standardized evaluation protocols [11], [12].

E. Toward Standardized Evaluation Protocols

Advancing sentence-level relevance modeling requires evaluation standards that extend beyond performance metrics:

- Semantic Faithfulness Testing: entailment-sensitive validation beyond lexical overlap [11], [20].
- Correlation Governance Auditing: detection of feature multicollinearity and proxy dominance [9], [21].
- Multi-Run Stability Reporting: confidence intervals and seed variance disclosure [12].
- Cross-Domain Robustness Testing: evaluation under distribution shift [6].
- Interpretability Validation: counterfactual or perturbation-based explanation verification [9].

Absent such standardized protocols, comparative claims remain performance-centric rather than governance-aware.

F. Open Challenges and Emerging Directions

The synthesis suggests several directions for advancing the field:

- Formalizing a unified multi-factor operational definition of relevance,
- Developing explicit inter-factor interaction models,
- Integrating correlation-aware regularization during signal construction,
- Establishing faithfulness-sensitive and entailment-aligned evaluation frameworks,
- Standardizing stability and reproducibility reporting,
- Designing governance protocols for LLM-based relevance scoring.

Addressing these challenges will shift sentence-level relevance modeling from performance-centric optimization toward explainable, correlation-governed, and reliability-sensitive frameworks.

VIII. CONCLUSION AND OUTLOOK

Sentence-level relevance modeling has evolved from heuristic lexical aggregation to relational graph centrality, neural scoring–selection architectures, transformer-based contextual encoding, and emerging LLM-assisted estimation. Across this progression, relevance has transitioned from a static, surface level similarity score to a context-conditioned, interaction aware construct shaped by semantic alignment, structural prominence, discourse dynamics, and redundancy constraints.

Formally, sentence-level relevance can be abstracted as a real-valued scoring function

$$R(si \mid C) = f(\phi(si, C)), \quad (5)$$

where si denotes a candidate sentence, C represents the conditioning context (e.g., document, query, task specification, or partial selection state), and ϕ denotes a multi-factor signal vector. In practice, ϕ integrates lexical, semantic, structural, discourse, and informativeness signals, while $f(\cdot)$ represents the fusion mechanism ranging from linear aggregation and graph propagation to learned neural composition and instruction-conditioned judgment. The induced ordering of R governs ranking and

selection decisions across retrieval, summarization, and reasoning tasks.

This survey highlighted three structural observations. First, relevance is inherently multi-factor and task-conditioned; static scoring assumptions are increasingly replaced by state-aware and utility-sensitive formulations. Second, relevance factors are often statistically correlated, and ungoverned fusion may lead to redundancy amplification, proxy-feature dominance, and unstable attribution. Correlation-aware modeling and representation stability therefore emerge as necessary complements to semantic expressiveness. Third, evaluation practices have not progressed proportionally with modeling sophistication. Overlap-based metrics provide incomplete proxies for informational validity, and stability across runs, domains, and prompt variations remains inconsistently reported.

Looking forward, advancing sentence-level relevance modeling requires unified operational definitions across tasks, explicit modeling of inter-factor interactions, correlation governed signal integration, and standardized stability and faithfulness-sensitive evaluation protocols. Such developments will enable relevance systems that are not only semantically powerful, but also interpretable, reproducible, and reliable under real-world deployment constraints and domain shift.

REFERENCES

[1] Y. Meena and D. Gopalani, "Analysis of sentence scoring methods for extractive automatic text summarization," in Proc. ICTCS, 2014.

[2] R. Ferreira et al., "Assessing sentence scoring techniques for extractive text summarization," Expert Systems with Applications, vol. 40, no. 14, pp. 5755–5764, 2013.

[3] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," Journal of Artificial Intelligence Research, 2004.

[4] R. Ferreira et al., "GETS: Sentence scoring scheme in graph-based extractive text summarization for text mining applications," Information Processing & Management, vol. 50, no. 5, pp. 851–866, 2014.

[5] Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao, "A joint sentence scoring and selection framework for neural extractive document summarization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 671–681, 2020.

[6] P. Baudiš et al., "Sentence pair scoring: Towards unified framework for text comprehension," arXiv preprint arXiv:1510.08398, 2016.

[7] K. Song, Y. Leng, X. Tan, Y. Zou, T. Qin, and D. Li, "Transcormer: Transformer for sentence scoring with sliding language modeling," arXiv preprint arXiv:2205.12986, 2022.

[8] "Recent advances in sentence-level relevance modeling," Information Processing & Management, 2023.

[9] E. AlJalaud and M. Hosny, "Enhancing explainable artificial intelligence using adaptive feature weight genetic explanation with Pearson correlation," Mathematics, vol. 12, no. 23, 2024.

[10] S. Ali et al., "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," Information Fusion, vol. 99, 2023.

[11] J. Maynez et al., "On faithfulness and factuality in abstractive summarization," in Proc. ACL, 2020.

[12] "Unified sentence scoring frameworks for relevance-based NLP applications," 2023.

[13] "Creating and evaluating multi-document sentence extract summaries," in Proc. ACM SIGIR, 2001.

[14] D. Patel, S. Shah, and H. Chhinkaniwala, "Fuzzy logic based multi-document summarization with improved sentence scoring and redundancy removal technique," Expert Systems with Applications, vol. 41, no. 12, pp. 5371–5383, 2014.

[15] J. P. Verma et al., "Graph-based extractive text summarization sentence scoring scheme for big data applications," Information, vol. 14, no. 9, 2023.

[16] J. Lin, X. Sun, S. Ma, and Q. Su, "Global encoding for abstractive summarization," in Proc. ACL, 2018.

[17] D. Jain, M. D. Borah, and A. Biswas, "A sentence is known by the company it keeps: Improving legal document summarization using deep clustering," Artificial Intelligence

and Law, vol. 32, no. 1, pp. 165–200, 2024.

[18] H. Wang et al., "Exploring explainable selection to control abstractive summarization," in Proc. AAAI, 2021.

[19] "An explainable framework for sentence-level relevance modeling," Information, 2023.

[20] A. Mishra et al., "Looking beyond sentence-level natural language inference for question answering and text summarization," in Proc. NAACL, 2021.

[21] Y. Liu et al., "Deep graph clustering via dual correlation reduction," arXiv preprint arXiv:2112.14772, 2021.