SIGIR 2022-2025:
The Evolution of AI and LLM Technologies for Search and Discovery

A Comprehensive Survey of Industry-Relevant Research

Research Compilation

October 2024

Abstract

This document provides a comprehensive survey of research from SIGIR conferences (2022-2025) documenting the evolution of artificial intelligence and large language model (LLM) applications in search and discovery systems. Spanning the critical period from pre-ChatGPT transformer-based retrieval (2022) through the explosion of LLM-augmented IR (2023-2025), this survey emphasizes practical, industry-applicable work in retrieval, ranking, query understanding, document understanding, RAG systems, and evaluation methodologies. The timeline captures the IR community's transformation: from BERT-based dense retrieval optimization (2022), through early LLM exploration for query generation and relevance feedback (2023), to mature LLM-integrated ranking systems and comprehensive RAG frameworks (2024-2025).

Contents

1 Introduction

The period from 2022 to 2025 represents a transformative era in information retrieval research, marking the transition from neural retrieval methods to LLM-augmented search systems. This survey tracks the evolution across four SIGIR conferences:

1.1 Timeline of Key Developments

- **SIGIR 2022** (Madrid, Spain, July 11-15, 2022): The transformer and BERT era
  - Foundation work in dense retrieval with pre-trained language models
  - Knowledge distillation for efficient neural rankers
  - Multi-stage ranking architectures becoming standard
- **SIGIR 2023** (Taipei, Taiwan, July 23-27, 2023): Early LLM exploration
  - ChatGPT released November 2022; first SIGIR post-ChatGPT
  - First Workshop on Generative Information Retrieval (Gen-IR)
  - LLMs for query generation, relevance feedback, and question answering
  - Sparse lexical representations with contextual embeddings
- **SIGIR 2024** (Washington, D.C., USA, July 14-18, 2024): LLM maturation
  - Dedicated Large Language Model Day
  - First Workshop on LLM Evaluation for IR (LLM4Eval)
  - Second Workshop on Generative IR
  - Scaling laws for dense retrieval established
  - RAG evaluation frameworks introduced
- **SIGIR 2025** (Padua, Italy, July 13-18, 2025): Production-ready LLM-IR
  - Robust RAG systems with collective intelligence
  - Efficiency optimizations for multi-vector retrieval
  - Zero-shot ranking with precomputed features
  - LLM-based generative recommendation systems

1.2 Survey Scope and Focus

This survey emphasizes **top-tier, peer-reviewed research with clear paths to production deployment**, covering:

- Dense retrieval evolution and scaling laws
- LLM-based ranking and reranking techniques
- Retrieval-Augmented Generation (RAG) systems and evaluation
- Generative retrieval and document representation

- Query understanding, reformulation, and expansion
- Evaluation metrics and methodologies for LLM-powered IR
- Efficiency optimizations for neural ranking systems

2 SIGIR 2022: Foundation Era

SIGIR 2022 occurred before the ChatGPT revolution but laid critical groundwork for LLM-augmented IR through transformer-based dense retrieval and neural ranking research.

2.1 Best Paper Awards

2.1.1 Best Paper: Non-Factoid Question Answering

Title: A Non-Factoid Question-Answering Taxonomy
Authors: Valeriia Bolotova, Vladislav Blinov, Falk Scholer, Bruce Croft, Mark Sanderson

Industry Relevance: Provides framework for understanding diverse question types beyond simple factoid QA.

Key Contributions:

- Comprehensive taxonomy of non-factoid questions (opinion, comparative, procedural)
- Evaluation frameworks for complex question answering systems
- Foundation for RAG systems handling diverse information needs

2.1.2 Best Short Paper: Curriculum Learning for Dense Retrieval Distillation

Authors: Hansi Zeng, Hamed Zamani, Vishwa Vinay
Institution: UMass Amherst CIIR, Adobe Research India
Code: github.com/HansiZeng/CL-DRD

Industry Relevance: Enables efficient dense retrieval models through knowledge distillation from expensive cross-encoders.

Key Contribution: CL-DRD Framework

- **Curriculum learning** for knowledge distillation: progressively increase training difficulty
- **Iterative optimization:** Start with coarse-grained preference pairs, move to fine-grained ordering
- **Teacher-student architecture:** Distill knowledge from cross-encoder (teacher) to bi-encoder (student)
- Strong performance on MS MARCO, TREC'19, TREC'20 benchmarks

Practical Implications:

- Reduces inference cost: bi-encoders 100-1000x faster than cross-encoders
- Maintains ranking quality while enabling real-time retrieval
- Applicable to any dense retrieval model architecture
- Foundation for modern multi-stage ranking pipelines

2.2 Key Research Themes

2.2.1 Pretrained Transformers for Text Ranking: BERT and Beyond

Tutorial at SIGIR 2021/2022
ArXiv: 2010.06467

Industry Relevance: Comprehensive guide to transformer-based ranking, still foundational in 2024.

Key Coverage:

- **Reranking architectures:** Cross-encoders for multi-stage pipelines
- **Dense retrieval:** Bi-encoders for first-stage candidate generation
- **Pre-training strategies:** BERT, RoBERTa, T5 for domain adaptation
- **Efficiency-effectiveness tradeoffs:** Late interaction models (ColBERT)

2.2.2 BERT-Based Dense Retrieval and Entity Ranking

Notable accepted papers:

- *BERT-based Dense Intra-ranking and Contextualized Late Interaction* (Minghan Li, Eric Gaussier)
- *BERT-ER: Query-Specific BERT Entity Representations* (Shubham Chatterjee, Laura Dietz)

Industry Impact:

- Established contextualized embeddings as standard for dense retrieval
- Entity-aware ranking for knowledge-intensive search (e.g., Wikipedia, enterprise search)
- Late interaction patterns balance effectiveness and efficiency

2.3 Production Insights from SIGIR 2022

Multi-Stage Ranking Architecture (2022 Best Practice):

1. **Stage 1:** BM25 or dense retrieval (bi-encoder) for candidate generation (top-1000)
2. **Stage 2:** Cross-encoder reranking (top-100)
3. **Optimization:** Use curriculum learning distillation (CL-DRD) to create efficient student models

Key Takeaway: SIGIR 2022 established transformer-based neural ranking as production-ready, setting the stage for LLM integration in subsequent years.

3 SIGIR 2023: Early LLM Era

SIGIR 2023 was the first conference after ChatGPT's November 2022 release, marking the beginning of systematic LLM exploration in information retrieval.

3.1 Best Paper Awards

3.1.1 Best Paper: IR Experiment Platform

Title: The Information Retrieval Experiment Platform
Authors: Maik Fröbe, Jan Heinrich Reimer, Sean MacAvaney, Niklas Deckers, Simon Reich, Janek Bevendorff, Benno Stein, Matthias Hagen, Martin Potthast

Industry Relevance: Infrastructure for reproducible IR experiments, critical for LLM-IR research validation.

Key Contributions:

- Standardized platform for IR experiment deployment and evaluation
- Reproducibility tooling for neural ranking experiments
- Foundation for comparing LLM-based vs. traditional IR approaches

3.1.2 Best Student Paper: Dense Retrieval for Visual QA

Title: A Symmetric Dual Encoding Dense Retrieval Framework for Knowledge-Intensive Visual Question Answering
Authors: Alireza Salemi, Juan Altmayer Pizzorno, Hamed Zamani
Institution: UMass Amherst

Industry Relevance: Multimodal retrieval for visual search and image-based Q&A systems.

Key Contributions:

- **Symmetric dual encoding:** Unified representation for images and text
- **Knowledge-intensive VQA:** Retrieval-augmented visual question answering
- Dense retrieval for multimodal contexts (precursor to vision-language models)

Note: Same lead author (Alireza Salemi) would win SIGIR 2024 Best Short Paper for RAG evaluation (eRAG), showing research trajectory toward LLM-augmented retrieval.

3.1.3 Best Short Paper: SparseEmbed

Title: SparseEmbed: Learning Sparse Lexical Representations with Contextual Embeddings for Retrieval
Authors: Weize Kong, Jeffrey M. Dudek, Cheng Li, Mingyang Zhang, Michael Bendersky

Industry Relevance: Bridges lexical and semantic retrieval, combining inverted index efficiency with neural understanding.

Key Contributions:

- **Sparse lexical representations:** Learned term importance weights (similar to SPLADE)
- **Contextual embeddings:** BERT-derived term representations

- **Inverted index compatibility:** Works with existing search infrastructure
- Hybrid approach: semantic understanding without dense embedding overhead

Practical Implications:

- Enables semantic search on traditional inverted indexes
- Lower latency than dense retrieval for exact-match queries
- Complementary to dense retrieval in hybrid systems

3.2 LLM-Specific Research at SIGIR 2023

3.2.1 First Workshop on Generative Information Retrieval (Gen-IR)

The inaugural Gen-IR workshop explored applying pre-trained generation models for IR tasks.

Key Themes:

- **Grounding challenges:** Ensuring generated answers are factually accurate
- **Attribution:** Citing sources for generated content
- **Bias mitigation:** Avoiding harmful or biased generations
- Early generative retrieval models (document ID generation)

3.2.2 Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?

Authors: S. Wang, H. Scells, B. Koopman, G. Zuccon

Industry Relevance: Academic and medical literature search, complex query formulation.

Key Findings:

- ChatGPT can generate reasonable Boolean queries from natural language descriptions
- Quality varies; requires expert validation for systematic reviews
- Useful for query bootstrapping and ideation

3.2.3 Can Generative LLMs Create Query Variants for Test Collections?

Authors: M. Alaofi et al.

Industry Relevance: Test collection creation, query augmentation, evaluation datasets.

Key Findings:

- LLMs can generate diverse query paraphrases
- Useful for creating evaluation datasets without manual annotation
- Foundation for synthetic query generation (precursor to DUQGen tutorial at SIGIR 2024)

3.2.4 Generative Relevance Feedback with Large Language Models

Authors: Iain Mackie, Shubham Chatterjee, Jeffrey Dalton
Institution: University of Edinburgh

Industry Relevance: Query reformulation, interactive search, conversational IR.

Key Contributions:

- **LLM-based relevance feedback:** Uses LLMs to reformulate queries based on initial results
- **Generative approach:** Creates new query terms rather than selecting from corpus
- Improves recall in interactive search sessions

Practical Implications:

- Reduces manual query refinement burden
- Applicable to conversational search systems
- Foundation for multi-turn retrieval dialogue

3.3 Other Notable SIGIR 2023 Papers

3.3.1 FiD-Light: Efficient Retrieval-Augmented Text Generation

Authors: Sebastian Hofstätter, Jiecao Chen, Karthik Raman, Hamed Zamani

Industry Relevance: Early work on efficient RAG systems.

Key Contributions:

- Optimizes Fusion-in-Decoder (FiD) architecture for efficiency
- Reduces computational overhead of multi-document generation
- Precursor to production RAG systems

3.3.2 Lexically-Accelerated Dense Retrieval

Authors: Hrishikesh Kulkarni, Sean MacAvaney, Nazli Goharian, Ophir Frieder

Industry Relevance: Hybrid retrieval combining lexical and dense approaches.

Key Contributions:

- Uses lexical signals to accelerate dense retrieval
- Reduces candidate set before dense scoring
- Practical hybrid architecture for production systems

3.4 Production Insights from SIGIR 2023

Emerging Best Practices:

- **Hybrid retrieval dominates:** Combine BM25, sparse embeddings (SparseEmbed/SPLADE), and dense retrieval
- **LLMs for query understanding:** ChatGPT and similar models useful for query generation and reformulation

- **RAG architectures emerging:** FiD-Light and similar work lay groundwork for production RAG
- **Evaluation challenges:** Generative IR requires new evaluation frameworks (addressed in SIGIR 2024)

4 SIGIR 2024: LLM Maturation Era

SIGIR 2024 marked the maturation of LLM-augmented IR with dedicated workshops, comprehensive evaluation frameworks, and production-ready systems.

4.1 Conference Highlights

- **Large Language Model Day:** Full day of LLM-focused presentations and panels
- **LLM4Eval Workshop:** 50+ attendees, focus on LLM-based evaluation for IR
- **Gen-IR Workshop (2nd edition):** Generative retrieval maturation
- **IR-RAG Workshop:** Information retrieval's role in RAG systems

4.2 Best Paper Awards

4.2.1 Best Long Paper: Scaling Laws For Dense Retrieval

Authors: Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, Yiqun Liu
Institutions: Tsinghua University, Renmin University of China
ArXiv: 2403.18684

Industry Relevance: Provides critical insights for capacity planning and ROI analysis in dense retrieval systems.

Key Contributions:

- First comprehensive study of scaling laws in dense retrieval
- Proposes **contrastive log-likelihood** as continuous evaluation metric
- Extensive experiments across model sizes (millions to billions of parameters)
- **Predictive framework:** Estimate performance gains before investing in larger models/datasets

Key Findings:

- Dense retrieval performance follows predictable scaling patterns (similar to language models)
- Data scaling and model scaling have different efficiency curves
- Contrastive log-likelihood correlates with downstream retrieval metrics

Practical Implications:

- Data-driven decisions: "Will doubling training data justify the cost?"
- Model size selection: Tradeoff between effectiveness and inference latency
- Benchmark for teams evaluating dense retrieval deployments

4.2.2 Best Long Paper: Workbench for Autograding RAG Systems

Author: Laura Dietz
Institution: University of New Hampshire

Industry Relevance: Critical tooling for teams deploying RAG systems in production.

Key Contributions:

- Automated evaluation framework for RAG pipelines
- Evaluates both retrieval quality and generation quality
- Production-ready tooling for CI/CD workflows

Practical Implications:

- Reduces manual evaluation burden in RAG development
- Enables A/B testing of retrieval and generation configurations
- Regression testing when updating components

4.2.3 Best Short Paper: Evaluating Retrieval Quality in RAG

Title: Evaluating Retrieval Quality in Retrieval-Augmented Generation
Authors: Alireza Salemi, Hamed Zamani
Institution: UMass Amherst
ArXiv: 2404.13781
Code: github.com/alirezasalemi7/eRAG

Industry Relevance: Directly addresses the evaluation challenge faced by every RAG deployment team.

Problem Statement:

- Traditional retrieval metrics (NDCG, MRR) show **weak correlation** with downstream RAG performance
- End-to-end RAG evaluation is computationally prohibitive

Key Contribution: eRAG Framework

- **Document-level evaluation:** Each retrieved document individually fed to LLM
- **Ground truth comparison:** Generated outputs evaluated against task labels
- **Relevance from performance:** Downstream performance becomes relevance label
- More efficient than full end-to-end evaluation

Practical Implications:

- Rapid iteration on retrieval without full RAG evaluation
- Identifies which retrieved documents contribute to generation quality
- Interpretable metrics for debugging RAG systems
- Open-source implementation accelerates adoption

4.2.4 Runner-up: Efficient Inverted Indexes for Learned Sparse Representations

Author: Sebastian Bruch et al.

Industry Relevance: Learned sparse representations (SPLADE) with traditional index efficiency.

Key Contributions:

- Optimized indexing structures for learned sparse models
- Semantic search with inverted index performance
- Enables hybrid retrieval architectures

4.2.5 Runner-up: A Reproducibility Study of PLAID

Authors: Sean MacAvaney, Nicola Tonellotto

Industry Relevance: PLAID (Performance-optimized Late Interaction Driver) for production dense retrieval.

Key Contributions:

- Validates PLAID's performance claims
- Deployment guidance and optimization insights
- Reproducibility best practices for dense retrieval research

4.3 LLM-Based Ranking and Reranking

4.3.1 Leveraging LLMs for Unsupervised Dense Retriever Ranking

Authors: Ekaterina Khramtsova, Shengyao Zhuang, Mahsa Baktashmotlagh, Guido Zuccon
Institution: University of Queensland

Industry Relevance: Addresses cold-start problem, reduces reliance on labeled data.

Key Contributions:

- **Unsupervised approach:** LLMs rank dense retrieval candidates without manual labels
- Combines dense retrieval (fast, semantic) with LLM reranking (accurate, nuanced)
- Applicable to new domains/languages with limited training data

4.3.2 A Setwise Approach for Zero-shot Ranking with LLMs

Industry Relevance: Zero-shot ranking enables immediate deployment.

Key Contributions:

- **Setwise ranking:** Multiple documents jointly rather than pairwise
- Reduces computational overhead vs. pairwise
- Balanced effectiveness-efficiency tradeoff

Ranking Architecture Comparison:

Approach	**Effectiveness**	**Efficiency**	**Use Case**
Pointwise	Moderate	High	First-stage retrieval
Pairwise	High	Low	Final reranking (top-k)
Setwise	High	Moderate	Mid-stage ranking
Listwise	Very High	Very Low	Offline evaluation

4.3.3 RLCF: Reinforcement Learning from Contrastive Feedback

Industry Relevance: Unsupervised alignment for domain-specific ranking.

Key Contributions:

- **Unsupervised alignment:** LLMs for IR tasks without human feedback
- Generates high-quality, context-specific responses
- Uses contrastive feedback from retrieval context

Practical Implications:

- Reduces dependency on expensive RLHF
- Continuous improvement from usage data
- Adapts LLMs to domain-specific ranking preferences

4.4 Retrieval-Augmented Generation (RAG)

4.4.1 CorpusLM: Unified Model for RAG and Generative Retrieval

Industry Relevance: Unifies multiple retrieval paradigms in single model.

Key Contributions:

- **Unified architecture** for three modes:
  - Generative retrieval (generates document IDs)
  - Closed-book generation (direct answer generation)
  - RAG (retrieve then generate)
- Single greedy decoding process for all modes
- Leverages external corpus for knowledge-intensive tasks

Practical Implications:

- Reduces model management overhead
- Runtime switching between retrieval strategies
- Potentially simpler deployment than multi-model RAG

4.4.2 Workshop: IR-RAG @ SIGIR 2024

The workshop acknowledged that despite RAG's prominence, systems require substantial improvement.

Key Themes:

- Retrieval quality is primary RAG bottleneck
- Traditional IR metrics poorly predict RAG performance (addressed by eRAG)
- Need for RAG-specific evaluation frameworks

4.5 Generative Retrieval and Document Representation

4.5.1 Workshop: Gen-IR @ SIGIR 2024 (2nd Edition)

Generative retrieval generates document identifiers autoregressively rather than retrieving from an index.

Core Challenge: Document Identifiers

Documents lack natural identifiers (unlike words in language modeling).

Identifier Strategies:

- **Unstructured atomic:** Random IDs (e.g., "doc_12345")
- **Semantically structured:** Hierarchical clustering-based IDs
- **Title-based:** Article titles as document IDs
- **URL-based:** Web URLs as natural identifiers
- **Term-sets:** Representative terms as identifiers

4.5.2 MERLIN: LLM-Generated Indices

Authors: Anirudh Ravichandran, Yidong Zou, Jayapragash Baskar, Anurag Beniwal

Key Contributions:

- Uses LLMs to generate semantically meaningful document identifiers
- Multiple enhanced representations for diverse query types
- Balances effectiveness and efficiency

4.5.3 Generative Retrieval as Multi-Vector Dense Retrieval

Key Insight: Generative retrieval with atomic identifiers is **equivalent** to single-vector dense retrieval. With hierarchical semantic identifiers, it behaves like hierarchical search in dense retrieval.

Industry Implications:

- Theoretical foundation for choosing between dense and generative retrieval
- Hierarchical identifiers enable interpretable retrieval paths
- Unifies understanding of dense and generative paradigms

4.6 Query Understanding and Reformulation

4.6.1 LDRE: Divergent Reasoning for Composed Image Retrieval

Authors: Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, Changsheng Xu

Industry Relevance: Multimodal query understanding for e-commerce, fashion, visual search.

Key Contributions:

- **Divergent reasoning:** Multiple interpretations of composed queries (image + text)
- **Ensemble approach:** Combines multiple reasoning paths
- **Zero-shot capability:** No training data for new query types

Use Cases:

- "Show me shoes like this but in red" (image + modification)
- "Find a sofa similar to this image but more modern"
- Visual + textual constraint search

4.6.2 Tutorial: DUQGen

Title: Effective Unsupervised Domain Adaptation of Neural Rankers by Diversifying Synthetic Query Generation

Key Contributions:

- **Diversified query generation:** Varied synthetic queries covering domain language
- **Unsupervised domain adaptation:** Adapt rankers to new domains without labels
- Uses LLMs to generate realistic queries from documents

Practical Workflow:

1. Generate diverse synthetic queries from domain documents using LLMs
2. Train neural ranker on synthetic (query, document) pairs
3. Optionally refine with small amounts of real user queries

4.7 Evaluation: LLM4Eval Workshop

The First Workshop on LLM Evaluation for IR (50+ attendees) focused on evaluation methodologies.

Focus Areas:

- LLM-based evaluation metrics for traditional IR
- LLM-based evaluation metrics for generative IR
- Effectiveness/efficiency of LLMs as relevance labelers
- Effectiveness/efficiency of LLMs as ranking models

4.7.1 Inter-Rater Reliability

Metrics:

- **Cohen's κ:** Inter-rater reliability (LLM vs. human)
- **Kendall's τ:** Agreement on system ranking order

Findings:

- Good agreement on system ordering (which systems perform better)
- More variation in exact relevance labeling (absolute scores)
- **Practical implication:** LLMs more reliable for **comparative evaluation** (A/B testing) than absolute judgments

4.7.2 Effectiveness-Efficiency Tradeoffs

Key Factors:

- **Model size:** Larger models (70B+) show better human agreement
- **Token consumption:** Longer context windows improve quality but increase cost
- **Latency:** Real-time evaluation requires optimized prompts
- **Prompt engineering:** Few-shot examples improve consistency

4.8 Dense Retrieval and Privacy

4.8.1 Vec2Text Privacy Threat to Dense Retrieval

Conference: SIGIR-AP 2024

Problem: Vec2Text can reconstruct original text from embeddings, raising privacy concerns for third-party embedding APIs (OpenAI, Cohere).

Key Contributions:

- Analysis of privacy risks in dense retrieval
- Mitigation strategies (embedding perturbation, differential privacy)
- Tradeoffs between privacy protection and retrieval quality

Practical Implications:

- Evaluate privacy risks before using third-party embedding services
- Consider self-hosted embedding models for sensitive data
- Implement protection mechanisms for regulated industries

4.9 Production Insights from SIGIR 2024

Recommended Multi-Stage Pipeline:

1. **Stage 1: Candidate Generation (Top-1000)**
   - Hybrid: BM25 + Dense retrieval
   - Optional: Learned sparse (SPLADE) for semantic understanding
   - Latency target: ¡50ms
2. **Stage 2: Neural Reranking (Top-100)**
   - Cross-encoder or pointwise LLM ranking
   - Latency target: ¡200ms
3. **Stage 3: LLM-Based Reranking (Top-10)**
   - Pairwise or setwise LLM ranking
   - Optional: RAG for answer generation
   - Latency target: ¡1s

RAG Best Practices:

- **Retrieval-first optimization:** Improve retrieval before generation tuning
- **Use eRAG framework:** Evaluate retrieval with downstream performance
- **Monitor separately:** Track retrieval metrics and generation quality independently
- **Self-refinement:** Implement iterative RAG with LLM feedback loops

5 SIGIR 2025: Production-Ready LLM-IR

SIGIR 2025 represents the maturation of LLM-augmented IR into production-ready systems with focus on robustness, efficiency, and real-world deployment.

5.1 Key Conference Themes

- **Robust RAG systems** with collective intelligence
- **Efficiency optimizations** for multi-vector retrieval
- **Zero-shot ranking** with precomputed features
- **LLM-based generative recommendation**
- **Fourth ReNeuIR Workshop:** Reaching efficiency in neural IR
- **First Robust-IR Workshop:** Robustness across domains and adversarial settings

5.2 Core RAG Research

5.2.1 CIRAG: Collective Intelligence RAG

Industry Relevance: Multi-agent RAG systems for complex reasoning.
Key Contributions:

- Integrates collective reasoning with retrieval
- Multiple LLM agents collaborate on retrieval-augmented tasks
- Improved performance on knowledge-intensive reasoning

5.2.2 Predicting RAG Performance for Text Completion

Industry Relevance: Capacity planning and performance prediction for RAG deployments.
Key Contributions:

- Framework for assessing RAG system performance before deployment
- Predictive models for RAG effectiveness based on retrieval quality
- Resource allocation guidance for production systems

5.2.3 Unveiling Knowledge Utilization in RAG

Industry Relevance: Understanding how LLMs leverage retrieved information.
Key Contributions:

- Analyzes how LLMs use retrieved documents
- Identifies which retrieved information influences generation
- Informs retrieval strategy optimization

5.2.4 Robust Fine-tuning for RAG against Retrieval Defects

Industry Relevance: Production RAG systems must handle imperfect retrieval.
Key Contributions:

- Improves resilience to noisy or irrelevant retrieved documents
- Fine-tuning strategies for robust RAG
- Handles retrieval failures gracefully

5.2.5 The Viability of Crowdsourcing for RAG Evaluation

Industry Relevance: Scalable human evaluation for RAG systems.
Key Contributions:

- Addresses human annotation challenges for RAG
- Crowdsourcing methodologies for RAG evaluation
- Cost-effective evaluation strategies

5.3 Query Understanding and LLM Limitations

5.3.1 LLM-based Query Expansion Fails for Unfamiliar and Ambiguous Queries

Industry Relevance: Identifies critical failure modes in LLM query processing.
Key Findings:

- LLMs struggle with unfamiliar domain terminology
- Ambiguous queries lead to inconsistent expansions
- Hybrid approaches (LLM + traditional query expansion) recommended

Practical Implications:

- Don't rely solely on LLMs for query understanding
- Combine LLM-based and traditional query expansion
- Domain-specific fine-tuning improves LLM query handling

5.3.2 Aligning Web Query Generation with Ranking Objectives

Industry Relevance: Optimizes synthetic query generation for ranking tasks.
Key Contributions:

- Uses Direct Preference Optimization (DPO) for query synthesis
- Aligns generated queries with ranking objectives
- Improves quality of synthetic training data

5.4 Dense and Sparse Retrieval Advances

5.4.1 On the Scaling of Robustness and Effectiveness in Dense Retrieval

Industry Relevance: Extends scaling laws (SIGIR 2024) to robustness analysis.
Key Contributions:

- Analyzes how model scaling affects robustness to domain shift
- Tradeoffs between effectiveness and robustness
- Guidance for model selection in production

5.4.2 IGP: Efficient Multi-Vector Retrieval via Proximity Graph Index

Industry Relevance: Optimizes late interaction models (ColBERT-style).
Key Contributions:

- Proximity graph indexing for multi-vector retrieval
- Reduces latency for ColBERT-style models
- Production-ready infrastructure for multi-vector search

5.4.3 WARP: Efficient Multi-Vector Retrieval Engine

Industry Relevance: Scalable multi-vector retrieval infrastructure.
Key Contributions:

- Efficient engine for multi-vector retrieval
- Handles large-scale deployments
- Optimized for ColBERT and similar late interaction models

5.5 Ranking and Reranking

5.5.1 Reason-to-Rank: Distilling LLM Reasoning for Reranking

Industry Relevance: Extracts LLM reasoning for efficient ranking.
Key Contributions:

- Distills direct and comparative reasoning from LLMs
- Creates efficient rankers from LLM knowledge
- Reduces inference cost while maintaining quality

5.5.2 Zero-Shot Reranking with Precomputed Features

Industry Relevance: Combines LLM capabilities with precomputed ranking features for efficiency.

Key Contributions:

- Hybrid approach: LLM zero-shot ranking + precomputed features
- Reduces LLM inference cost
- Maintains effectiveness through feature integration

Practical Implications:

- Enables LLM ranking at scale
- Precompute features offline, use LLM for final scoring
- Balances cost and quality

5.5.3 Efficient Re-ranking via Early Exit

Industry Relevance: Accelerates cross-encoder reranking.

Key Contributions:

- Early termination for confident predictions
- Reduces average inference time for cross-encoders
- Maintains ranking quality

5.6 Generative Retrieval and Recommendation

5.6.1 Information Retrieval in the Age of Generative AI: The RGB Model

Industry Relevance: Proposes new framework for generative retrieval.

Key Contributions:

- RGB (Retrieve, Generate, Blend) model
- Unifies retrieval and generation paradigms
- Theoretical framework for generative IR systems

5.6.2 Constrained Auto-Regressive Decoding in Generative Retrieval

Industry Relevance: Analyzes limitations of autoregressive generation for retrieval.

Key Findings:

- Autoregressive decoding constraints limit generative retrieval effectiveness
- Beam search tradeoffs between diversity and precision
- Guidance for generative retrieval system design

5.6.3 Order-agnostic Identifier for LLM-based Generative Recommendation

Industry Relevance: Recommendation systems using generative models.

Key Contributions:

- Position-invariant identifiers for item generation
- Reduces ordering bias in generative recommendation
- Applicable to product recommendation, content recommendation

5.7 Efficiency and Robustness Workshops

5.7.1 ReNeuIR 2025: Fourth Workshop on Efficiency

Focus: Holistic evaluation of neural IR methods including computational cost.

Key Themes:

- Efficiency benchmarks for modern IR systems
- Shared task on efficiency-oriented IR
- Training and inference optimization strategies
- Environmental impact of neural IR models

5.7.2 Robust-IR 2025: First Workshop on Robustness

Focus: Robustness across domains, adversarial settings, and distribution shifts.

Key Themes:

- Domain adaptation for neural rankers
- Adversarial robustness in retrieval
- Out-of-distribution generalization
- Robustness evaluation methodologies

5.8 Production Insights from SIGIR 2025

RAG System Architecture (2025 Best Practice):

- **Multi-agent RAG:** CIRAG-style collective intelligence for complex reasoning
- **Robust retrieval:** Fine-tune for resilience to retrieval defects
- **Predictive performance:** Use frameworks to predict RAG effectiveness before deployment
- **Knowledge utilization analysis:** Monitor which retrieved information influences generation

Ranking Optimization:

- **Zero-shot with features:** Combine LLM zero-shot ranking with precomputed features
- **Early exit reranking:** Accelerate cross-encoders with confidence-based termination

- **Distilled LLM reasoning:** Extract reasoning from LLMs into efficient rankers (Reason-to-Rank)

Multi-Vector Retrieval:

- **Production infrastructure:** IGP and WARP for scalable ColBERT-style retrieval
- **Proximity graph indexing:** Optimize multi-vector search latency
- **Late interaction models:** Balance effectiveness (cross-encoder quality) with efficiency (bi-encoder speed)

Query Understanding:

- **Hybrid query expansion:** Combine LLM-based and traditional methods (LLMs fail on unfamiliar/ambiguous queries)
- **Alignment with ranking:** Use DPO to align query generation with ranking objectives
- **Domain adaptation:** Fine-tune LLMs for domain-specific query understanding

6 Cross-Year Analysis and Evolution

6.1 Key Technology Transitions

Year	Dominant Technology	Key Innovation
2022	BERT-based dense retrieval	Curriculum learning distillation (CL-DRD) for efficient bi-encoders
2023	Early LLM integration	Generative relevance feedback, sparse lexical embeddings (SparseEmbed)
2024	LLM-augmented ranking	Scaling laws, RAG evaluation (eRAG), LLM-based ranking
2025	Production LLM-IR	Robust RAG, efficient multi-vector retrieval, zero-shot with features

6.2 Retrieval Architecture Evolution

2022 Standard Architecture:

1. BM25 or dense bi-encoder (candidate generation)
2. Cross-encoder reranking
3. Optional: Knowledge distillation for efficiency (CL-DRD)

2023 Emerging Architecture:

1. Hybrid retrieval: BM25 + dense + sparse embeddings (SparseEmbed/SPLADE)
2. Cross-encoder or early LLM reranking
3. LLM-based query reformulation

2024 Best Practice Architecture:

1. Multi-stage hybrid: BM25 + dense + learned sparse
2. Pointwise/setwise LLM ranking (mid-stage)
3. Pairwise LLM reranking (final stage)
4. RAG for answer generation (optional)
5. Evaluation via eRAG framework

2025 Production Architecture:

1. Efficient multi-vector retrieval (WARP/IGP)

2. Zero-shot LLM ranking with precomputed features
3. Early exit cross-encoder reranking
4. Robust multi-agent RAG (CIRAG)
5. Continuous robustness monitoring

6.3 RAG Evolution Timeline

- **2022:** Foundational work (non-factoid QA taxonomy, efficient generation)
- **2023:** Early RAG systems (FiD-Light), generative IR workshop
- **2024:** RAG evaluation frameworks (eRAG), CorpusLM unified model, IR-RAG workshop
- **2025:** Production-ready robust RAG (CIRAG), performance prediction, resilience to retrieval defects

6.4 Evaluation Methodology Evolution

- **2022:** Traditional IR metrics (NDCG, MRR), reproducibility studies
- **2023:** Early generative IR evaluation challenges identified
- **2024:** LLM4Eval workshop, eRAG framework, LLM-based evaluation metrics, inter-rater reliability studies
- **2025:** Crowdsourcing for RAG evaluation, robustness evaluation, efficiency benchmarks (ReNeuIR)

6.5 Query Understanding Evolution

- **2022:** Traditional query expansion, entity-aware ranking
- **2023:** LLM-based query generation (ChatGPT for Boolean queries), generative relevance feedback
- **2024:** Synthetic query generation for domain adaptation (DUQGen), multimodal query understanding (LDRE)
- **2025:** Limitations identified (LLMs fail on unfamiliar/ambiguous queries), alignment with ranking objectives (DPO)

7 Practical Implementation Guide

7.1 Choosing the Right Retrieval Architecture

7.1.1 When to Use Dense Retrieval

- **Use case:** Semantic search, concept matching, cross-lingual retrieval
- **Best for:** Natural language queries, synonym matching
- **2025 recommendation:** Combine with BM25 for hybrid approach
- **Infrastructure:** WARP/IGP for multi-vector (ColBERT-style)

7.1.2 When to Use Learned Sparse Retrieval

- **Use case:** Balance semantic understanding with inverted index efficiency
- **Best for:** Large-scale systems requiring both semantic and lexical matching
- **Technology:** SparseEmbed (SIGIR 2023), SPLADE
- **Advantage:** Works with existing search infrastructure

7.1.3 When to Use Generative Retrieval

- **Use case:** Experimental systems, small corpora, URL/title-based retrieval
- **Best for:** Human-interpretable identifiers, hierarchical retrieval
- **Limitation:** Constrained autoregressive decoding (SIGIR 2025 findings)
- **Maturity:** Emerging technology, not yet production-standard for large scale

7.2 LLM Ranking: Effectiveness vs. Cost

Approach	Quality	Latency	Cost	Recommended Stage
Pointwise	++	+	+	Stage 2 (top-100)
Setwise	+++	++	++	Stage 2-3 (top-50)
Pairwise	++++	+++	+++	Stage 3 (top-10)
Listwise	+++++	++++	++++	Offline only
Zero-shot + Features	+++	++	+	Stage 2 (SIGIR 2025)

Key: + (low) to +++++ (very high)

7.3 RAG System Design Checklist

1. **Retrieval optimization** (most important, per IR-RAG workshop):
   - Hybrid retrieval (BM25 + dense + learned sparse)
   - Evaluate with eRAG framework (SIGIR 2024)
   - Monitor retrieval quality independently from generation
2. **Robustness** (SIGIR 2025):
   - Fine-tune for resilience to retrieval defects
   - Handle noisy or irrelevant documents gracefully
   - Implement fallback strategies
3. **Multi-agent architecture** (optional, SIGIR 2025):
   - CIRAG-style collective intelligence for complex reasoning
   - Multiple LLM agents for verification and cross-checking
4. **Evaluation**:

- Use eRAG for retrieval component evaluation
- Separate metrics for retrieval and generation quality
- Crowdsourcing for large-scale human evaluation (SIGIR 2025)

5. **Performance prediction** (SIGIR 2025):
   - Use frameworks to predict RAG effectiveness before deployment
   - Capacity planning based on retrieval quality predictions

7.4 Evaluation Strategy

7.4.1 Offline Evaluation

- **Retrieval:** Traditional IR metrics (NDCG, MRR) + eRAG downstream performance
- **Ranking:** LLM-based evaluation (pointwise for speed, pairwise for accuracy)
- **RAG:** eRAG framework, autograding workbench (SIGIR 2024)
- **Agreement metrics:** Kendall's τ for system ranking, Cohen's κ for labeling

7.4.2 Online Evaluation

- A/B testing with user engagement metrics
- Click-through rate, dwell time, session success
- Continuous monitoring of robustness (domain shift, adversarial queries)

7.5 Scaling Considerations

Data vs. Model Size (from SIGIR 2024 scaling laws):

- Use contrastive log-likelihood for continuous scaling analysis
- Quantify expected performance gains before scaling investments
- ROI analysis: doubling training data vs. doubling model size
- Diminishing returns at high scales

Robustness vs. Effectiveness (SIGIR 2025):

- Larger models more robust to domain shift
- Tradeoff: effectiveness improvements vs. robustness gains
- Consider deployment domain when selecting model size

8 Future Directions and Open Challenges

8.1 Emerging Trends (2025 and Beyond)

- **Unified retrieval models:** Single models handling RAG, generative retrieval, and closed-book generation (CorpusLM trend)
- **Multi-agent RAG:** Collective intelligence systems (CIRAG) for complex reasoning
- **Efficient multi-vector retrieval:** Production infrastructure (WARP, IGP) enabling ColBERT-style models at scale
- **Zero-shot ranking with features:** Combining LLM capabilities with precomputed signals for cost-effective ranking
- **Robustness standardization:** Community-wide benchmarks for domain adaptation, adversarial robustness, OOD generalization
- **Privacy-preserving embeddings:** Addressing Vec2Text-style attacks on dense retrieval
- **Environmental impact:** Efficiency benchmarks (ReNeuIR) and environmental considerations

8.2 Open Challenges

8.2.1 Generative Retrieval

- **Document identifier design:** No consensus on optimal identifier strategy
- **Scaling to large corpora:** Constrained autoregressive decoding limits effectiveness (SIGIR 2025)
- **Index updates:** Handling dynamic document collections

8.2.2 RAG Systems

- **Retrieval quality bottleneck:** Still the primary limitation (IR-RAG workshop consensus)
- **Robustness:** Handling noisy, irrelevant, or contradictory retrieved documents
- **Evaluation:** Need for standardized RAG benchmarks beyond eRAG
- **Knowledge utilization:** Better understanding of how LLMs use retrieved information

8.2.3 LLM-Based Ranking

- **Cost-effectiveness:** Balancing quality with inference costs
- **Latency:** Meeting real-time ranking requirements (¡100ms)
- **Calibration:** LLMs poorly calibrated for relevance judgments
- **Bias:** Position bias, length bias, popularity bias in LLM rankings

8.2.4 Query Understanding

- **LLM limitations:** Failure on unfamiliar and ambiguous queries (SIGIR 2025)
- **Domain adaptation:** Expensive fine-tuning for domain-specific terminology
- **Multimodal queries:** Limited work on complex visual + text queries

8.2.5 Evaluation

- **LLM evaluation reliability:** Better for comparative than absolute judgments
- **RAG-specific metrics:** Beyond eRAG, need comprehensive evaluation frameworks
- **Efficiency benchmarks:** Standardized evaluation including computational cost
- **Robustness evaluation:** Methodologies for adversarial and OOD settings

8.3 Research Priorities for Industry

1. **Hybrid retrieval optimization:** Finding optimal combination of BM25, dense, and learned sparse
2. **Efficient LLM ranking:** Reducing cost through distillation (Reason-to-Rank), early exit, feature integration
3. **Robust RAG:** Handling retrieval defects, improving resilience, multi-agent verification
4. **Multi-vector retrieval infrastructure:** Production-ready systems for ColBERT-style models
5. **Domain adaptation:** Cost-effective methods for adapting to new domains (synthetic query generation, few-shot learning)
6. **Privacy and security:** Protecting embeddings, secure retrieval APIs, differential privacy
7. **Evaluation tooling:** Automated, scalable evaluation for rapid iteration

9 Conclusion

The period from SIGIR 2022 to 2025 captures the complete transformation of information retrieval from neural ranking optimization to mature LLM-augmented search systems. Key insights for practitioners:

9.1 Core Lessons

1. **Hybrid retrieval is essential** (2022-2025 consensus):
   - Combine BM25 (precision), dense retrieval (semantic), learned sparse (hybrid benefits)
   - No single approach dominates all scenarios
2. **Multi-stage ranking balances effectiveness and efficiency**:
   - Fast candidate generation (top-1000): hybrid retrieval

- Mid-stage reranking (top-100): cross-encoder or pointwise LLM
- Final reranking (top-10): pairwise LLM with high-quality scoring

3. **RAG requires retrieval-first optimization**:
   - Retrieval quality is the primary bottleneck (IR-RAG workshop)
   - Use eRAG framework for retrieval evaluation (SIGIR 2024)
   - Fine-tune for robustness to retrieval defects (SIGIR 2025)
4. **Scaling laws enable predictability** (SIGIR 2024):
   - Data-driven decisions about model size and training data
   - Contrastive log-likelihood for continuous scaling analysis
   - ROI analysis before major investments
5. **LLM ranking requires careful cost management**:
   - Use pointwise for speed, pairwise for accuracy
   - Zero-shot with precomputed features (SIGIR 2025)
   - Distill LLM reasoning into efficient models (Reason-to-Rank)
6. **Evaluation must evolve with technology**:
   - Traditional IR metrics insufficient for RAG (use eRAG)
   - LLMs better for comparative than absolute evaluation
   - Monitor retrieval and generation quality separately
7. **Robustness is critical for production**:
   - Domain adaptation challenges remain (SIGIR 2025)
   - Handle retrieval defects gracefully
   - Monitor for adversarial queries and distribution shift
8. **Efficiency considerations are paramount**:
   - ReNeuIR workshop series (2023-2025) emphasizes computational cost
   - Environmental impact of neural IR models
   - Tradeoffs between effectiveness, efficiency, and robustness

9.2 Technology Maturity Assessment

Technology	Maturity	Recommendation
Dense retrieval	Production-ready	Deploy with hybrid approach
Learned sparse	Production-ready	Deploy for semantic + lexical
LLM reranking	Maturing	Deploy with cost controls
RAG systems	Maturing	Deploy with robust retrieval
Generative retrieval	Experimental	Research only, not production
Multi-agent RAG	Emerging	Pilot for complex reasoning

9.3 Final Recommendations

For practitioners building search and discovery systems in 2025 and beyond:

1. **Start with hybrid retrieval:** BM25 + dense + learned sparse (SparseEmbed/SPLADE)
2. **Use multi-stage ranking:** Invest in fast candidate generation, reserve expensive LLM ranking for final stage
3. **Optimize retrieval before generation:** In RAG systems, focus on retrieval quality first
4. **Leverage scaling laws:** Use predictive frameworks (SIGIR 2024) for capacity planning
5. **Implement robust evaluation:** eRAG for RAG systems, LLM-based evaluation for comparative analysis
6. **Plan for efficiency:** Use distillation (CL-DRD), early exit, zero-shot with features to reduce costs
7. **Monitor robustness:** Test across domains, handle retrieval defects, watch for adversarial queries
8. **Stay hybrid:** Don't replace traditional IR wholesale; combine strengths of lexical, dense, and LLM approaches

The research from SIGIR 2022-2025 provides a robust foundation for building production-grade search and discovery systems in the age of large language models. Success requires understanding the strengths and limitations of each approach, combining techniques strategically, and maintaining focus on the fundamentals: retrieval quality, efficiency, and robustness.

10 References

10.1 SIGIR 2022

- Valeriia Bolotova et al., "A Non-Factoid Question-Answering Taxonomy," *SIGIR '22*, July 2022. (Best Paper)
- Hansi Zeng, Hamed Zamani, Vishwa Vinay, "Curriculum Learning for Dense Retrieval Distillation," *SIGIR '22*, July 2022. (Best Short Paper) Code: `https://github.com/HansiZeng/CL-DRD`
- Andrew Yates et al., "Pretrained Transformers for Text Ranking: BERT and Beyond," ArXiv: 2010.06467, Tutorial.
- Minghan Li, Eric Gaussier, "BERT-based Dense Intra-ranking and Contextualized Late Interaction," *SIGIR '22*, July 2022.
- Shubham Chatterjee, Laura Dietz, "BERT-ER: Query-Specific BERT Entity Representations," *SIGIR '22*, July 2022.

10.2 SIGIR 2023

- Maik Fröbe et al., "The Information Retrieval Experiment Platform," *SIGIR '23*, July 2023. (Best Paper)
- Alireza Salemi, Juan Altmayer Pizzorno, Hamed Zamani, "A Symmetric Dual Encoding Dense Retrieval Framework for Knowledge-Intensive Visual Question Answering," *SIGIR '23*, July 2023. (Best Student Paper)
- Weize Kong et al., "SparseEmbed: Learning Sparse Lexical Representations with Contextual Embeddings for Retrieval," *SIGIR '23*, July 2023. (Best Short Paper)
- S. Wang, H. Scells, B. Koopman, G. Zuccon, "Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?" *SIGIR '23*, July 2023.
- M. Alaofi et al., "Can Generative LLMs Create Query Variants for Test Collections?" *SIGIR '23*, July 2023. (Short paper)
- Iain Mackie, Shubham Chatterjee, Jeffrey Dalton, "Generative Relevance Feedback with Large Language Models," *SIGIR '23*, July 2023. (Short paper)
- Sebastian Hofstätter et al., "FiD-Light: Efficient Retrieval-Augmented Text Generation," *SIGIR '23*, July 2023.
- Hrishikesh Kulkarni et al., "Lexically-Accelerated Dense Retrieval," *SIGIR '23*, July 2023.
- "Gen-IR @ SIGIR 2023: The First Workshop on Generative Information Retrieval," July 2023.

10.3 SIGIR 2024

- Yan Fang et al., "Scaling Laws For Dense Retrieval," *SIGIR '24*, July 2024. ArXiv: 2403.18684. (Best Long Paper)
- Laura Dietz, "A Workbench for Autograding Retrieve/Generate Systems," *SIGIR '24*, July 2024. (Best Long Paper)
- Alireza Salemi, Hamed Zamani, "Evaluating Retrieval Quality in Retrieval-Augmented Generation," *SIGIR '24*, July 2024. ArXiv: 2404.13781. Code: https://github.com/alirezasalemi7/eRAG (Best Short Paper)
- Sebastian Bruch et al., "Efficient Inverted Indexes for Approximate Retrieval over Learned Sparse Representations," *SIGIR '24*, July 2024. (Runner-up)
- Sean MacAvaney, Nicola Tonellotto, "A Reproducibility Study of PLAID," *SIGIR '24*, July 2024. (Runner-up)
- Ekaterina Khramtsova et al., "Leveraging LLMs for Unsupervised Dense Retriever Ranking," *SIGIR '24*, July 2024.
- "A Setwise Approach for Effective and Highly Efficient Zero-shot Ranking with Large Language Models," *SIGIR '24*, July 2024.
- "RLCF: Reinforcement Learning from Contrastive Feedback," *SIGIR '24*, July 2024.
- "CorpusLM: Unified Language Model using RAG and Generative Retrieval," *SIGIR '24*, July 2024.
- Anirudh Ravichandran et al., "MERLIN: Multiple enhanced representations with LLM generated indices," *SIGIR '24*, July 2024.
- "Generative Retrieval as Multi-Vector Dense Retrieval," *SIGIR '24*, July 2024.
- Zhenyu Yang et al., "LDRE: LLM-based Divergent Reasoning and Ensemble for Zero-Shot Composed Image Retrieval," *SIGIR '24*, July 2024.
- "DUQGen: Effective Unsupervised Domain Adaptation of Neural Rankers," Tutorial, *SIGIR '24*, July 2024.
- "Understanding and Mitigating the Threat of Vec2Text to Dense Retrieval Systems," *SIGIR-AP 2024*.
- "Report on the 1st Workshop on LLM for Evaluation in IR (LLM4Eval 2024)," ArXiv: 2408.05388, August 2024.
- "Gen-IR @ SIGIR 2024: The Second Workshop on Generative Information Retrieval," July 2024.
- "IR-RAG @ SIGIR 2024: Information Retrieval's Role in RAG Systems," Workshop, July 2024.
- "Recent Advances in Generative Information Retrieval," Tutorial, *SIGIR '24*, July 2024.
- "Robust Information Retrieval," Tutorial, *SIGIR '24*, July 2024.
- Zhai, "Large Language Models and Future of Information Retrieval: Opportunities and Challenges," Keynote, *SIGIR '24*, July 2024.

10.4 SIGIR 2025

- "CIRAG: Retrieval-Augmented Language Model with Collective Intelligence," *SIGIR '25*, July 2025.
- "Predicting RAG Performance for Text Completion," *SIGIR '25*, July 2025.
- "Unveiling Knowledge Utilization Mechanisms in LLM-based RAG," *SIGIR '25*, July 2025.
- "Robust Fine-tuning for RAG against Retrieval Defects," *SIGIR '25*, July 2025.
- "The Viability of Crowdsourcing for RAG Evaluation," *SIGIR '25*, July 2025.
- "LLM-based Query Expansion Fails for Unfamiliar and Ambiguous Queries," *SIGIR '25*, July 2025.
- "Aligning Web Query Generation with Ranking Objectives via DPO," *SIGIR '25*, July 2025.
- "On the Scaling of Robustness and Effectiveness in Dense Retrieval," *SIGIR '25*, July 2025.
- "IGP: Efficient Multi-Vector Retrieval via Proximity Graph Index," *SIGIR '25*, July 2025.
- "WARP: An Efficient Engine for Multi-Vector Retrieval," *SIGIR '25*, July 2025.
- "Reason-to-Rank: Distilling LLM Reasoning for Document Reranking," *SIGIR '25*, July 2025.
- "Zero-Shot Reranking with LLMs and Precomputed Ranking Features," *SIGIR '25*, July 2025.
- "Efficient Re-ranking with Cross-encoders via Early Exit," *SIGIR '25*, July 2025.
- "Information Retrieval in the Age of Generative AI: The RGB Model," *SIGIR '25*, July 2025.
- "Constrained Auto-Regressive Decoding Constrains Generative Retrieval," *SIGIR '25*, July 2025.
- "Order-agnostic Identifier for LLM-based Generative Recommendation," *SIGIR '25*, July 2025.
- "ReNeuIR at SIGIR 2025: The Fourth Workshop on Reaching Efficiency in Neural IR," July 2025.
- "Robust-IR @ SIGIR 2025: The First Workshop on Robust IR," ArXiv: 2503.18426, July 2025.

10.5 Conference Information

- **SIGIR 2022:** Madrid, Spain, July 11-15, 2022. https://sigir.org/sigir2022/
- **SIGIR 2023:** Taipei, Taiwan, July 23-27, 2023. https://sigir.org/sigir2023/
- **SIGIR 2024:** Washington, D.C., USA, July 14-18, 2024. https://sigir-2024.github.io/
- **SIGIR 2025:** Padua, Italy, July 13-18, 2025. https://sigir2025.dei.unipd.it/