A Survey on Semantic Searching and the Analysis of Long-Form Narrative Fiction [E]

David Liang
dlian4@illinois.edu
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Abstract

Recent advancements in Large Language Models (LLMs) have show remarkable semantic understanding across a range of tasks [15]. While they have a well-established success in summarizing content and short semantic understanding, applying these models to long-form narrative fiction, such as Chinese web novels with possibly thousands of chapters and millions of words, remains a significant challenge. These words often exceed even the huge context windows of modern LLMs, which additionally are burdened by performance degradation over long sequences [23], diluted attention, and difficulties in maintaining consistent understandings of entities or narrative structures across large corpora.

This paper is a survey of methods aimed at addressing these limitations, which were primarily aggregated for insights during the development of INK (Infrastructure for Narrative Knowledge), a personal project which is a conceptual framework for semantically working with long-form fiction. INK enables various abstractions that helps with things like plot-based semantic searching, automatic multi-label tagging (including both surface-level like #fighting, deep structural ones like #non-linear-timeline, and even meta ones like #adapted-to-movie), entity tracking across long spans, and multilingual analysis.

We explore techniques that help support these goals, including how to represent long documents (e.g., summarization [12, 20]), instruction-tuned embeddings [11, 34], Retrieval-Augmented Generation (RAG) and its extensions (e.g., FLARE [19], GraphRAG [8]), and methods for plot retrieval [17, 33]. Additionally, we look at a few approaches to Extreme Multi-label Text Classification (XMTC) [6, 43, 46] and the critical role of zero- and few-shot learning [2, 3], given the impracticality of training models for every narrative or analytical task provided the task fits in the context window. Finally, we look at some ongoing challenges in scalability, representation, evaluation [4, 41], and temporal modeling [18, 21], looking for a possible solution towards more robust, automated analysis of long-form fiction.

1 Introduction

In the world of online fiction, long-form narratives like Chinese web novels have produced massive multi-lingual digital corpora that require handling beyond traditional keyword searching. These stories, which are often serialized and span millions of words, require systems that can understand theme, plot, and structure beyond surface-level terms to assist in their recommendation or search. Semantic search offers a solution by retrieving based on intent and meaning, not just keyword overlap [14], which is a requirement for navigating these dense narrative texts. However, the length and complexity of these works pose serious challenges. Most large language models (LLMs) have fixed context windows that fall far short of an entire novel [23], and even those with longer contexts face performance and cost trade-offs [32]. Worse still, high-context models can struggle with cross-novel comparison or fail to consistently track entities and plot threads. This motivates the creation of our personal project framework called INK (Infrastructure for Narrative Knowledge), designed to handle, scale, and support tasks like plot-based search, tagging (e.g., via XMTC [6]), translation consistency [5], and multilingual alignment across massive narrative datasets. Fine-tuning for every task is infeasible, which is why zero-shot and few-shot methods [2, 3], methods that leverage pre-trained LLMs without additional training, are especially valuable. This survey focuses on approaches that use LLMs as tools, not as endpoints: instruction-tuned models, structured retrieval (e.g., RAG [15], FLARE [19]), tagging via semantic multi-label classification, and models that support narrative structure and temporal modeling. We explore how these systems manage length, semantic richness, and plot complexity at scale while outlining a few key gaps and potential paths forward.

2 Representing Long-Form Narrative Fiction

Effective semantic search over long-form narratives requires a method of representing the stories. Novels are not just long documents, but complex narrative systems with temporal structure, recurring entities, and implicit meaning which all have even more complex inter- and intra-document dependencies.

2.1 Foundational Concepts: Narrative Structure as Data

Narratives operate on multiple levels. There is the surface text, consisting of the raw words written. There is the sequence of events which allow one setting to flow to the next, one state to transition to the next. There are many character relationships as well as temporal dynamics. Computationally, that means that we have to track entities, actions, and the flow of time [9, 17]. A key framing comes from the classic distinction between *fabula* (underlying event order) and *syuzhet* (presented order) [17]. Plot-based search should favor "what happened" over "how it was told". While classic models (Freytag, Propp) provide useful structure, they don't scale well across genres or languages. Greimas' Actantial Model offers a more functional view by abstracting roles (Subject, Helper, Opponent, etc.), and recent work shows that LLMs can extract these actants reliably [9]. The open question is how to translate this theory into scalable, automated representations across arbitrarily huge numbers of books.

2.2 Embedding Long Documents

Most embedding models break under long-form narrative content. The usual, such as BERT or RoBERTa, have a context length of just a few hundred tokens. There are longer models, but then cost scales very, very quickly [32, 47]. There are roughly three options forward:

- **Chunk and Pool**: Divide the document, embed chunks, aggregate the result [47]. This is a simple method, but it risks breaking narrative flow and losing important long-range dependencies. Also, the chunking strategy itself becomes a critical design decision [13].
- **Native Long-Context Models**: These models (e.g., LongT5, Claude3.5, GPT-4o, Gemini 2.5) push the context window to 8k-128k-1M tokens, but at a significant computational price [32, 47].
- **Training-Free Context Extension**: These methods retrofit existing models to handle longer contexts without retraining. Position Interpolation and NTK-Aware RoPE [47] methods can extend context windows to 32k+ with minimal cost. Benchmarks like LONGEMBED [47] show the tradeoffs (there is better reach, but also a potential weaker deep understanding compared to actual trained long-sequence models).

Choosing between these options depends on whether the task at hand is optimizing for throughput, accuracy, or structural fidelity.

2.3 Narrative-Aware and Task-Specific Embeddings

Off-the-shelf embeddings usually miss the minute structural differences in fiction. Several methods exist to try and fix this:

- **StoryEmb**: Trained specifically to capture plot similarity across multiple summaries. Strong at retrieving alternative versions of the same story (e.g., remakes) [17].
- **Summarization-based Representations**: Extractive methods like Ranksum [20] or retrieval-guided models like SARESG [12] use condensed text as proxies for plot. Hierarchical summarization helps recover high-level structure. *(Outside of this survey, when applied to INK, this resulted in a method to perform plot searching by generating pseudo-documents of the query and the documents and using a basic hybrid BM25 and Vector Embedding similarity search to rank them)*
- **INSTRUCTOR**: Instruction-tuned models that generate task-aware embeddings depending on the prompt [11, 34]. They adapt surprisingly well to narrative use-cases (e.g., "embed for plot similarity").
- **HEAL**: This method aligns documents with an existing topic hierarchy and uses contrastive learning. Potentially useful for modeling chapters, arcs, or acts in novels [1].
- **Greimas Embeddings**: Apply LLMs to extract Greimas roles, then embed them separately [9]. This adds structure-awareness and helps distinguish documents with similar topics but different functional dynamics.

A combination of something that provides narrative alignment (StoryEmb, Greimas) with instruction-based flexibility (INSTRUCTOR) seems promising *and will likely be the next thing I incorporate into INK*.

2.4 Temporal Embeddings and Narrative Progression

Stories follow a series of events, i.e., time. Static embeddings are unable to capture this. Temporal models try to capture narrative change, like how characters evolve, how themes shift. Various approaches include:

- **Dynamic Contextualized Word Embeddings (DCWEs)**: Words get time-sensitive embeddings that reflect how their meaning changes with context or narrative phase [18].
- **Slice-Based Entity Embeddings**: Learn how character representations evolve across narrative "slices" (e.g., chapters), then compare trajectories [21]. This can reveal character arcs or plot pivots.
- **Time-Series Modeling**: Treat text as a sequence aligned with external or internal time. TaTS, TWM, and CTRM explore this in various forms [7, 27, 30].
- **Interaction-Based Embeddings (e.g., JODIE)**: Borrowed from recommendation systems, these track embeddings that evolve based on interactions (e.g., dialogue, relationships) [22].

Temporal modeling is still in its early-stage, but critical as stories cannot be understood as bags of sentences with no order. Embeddings need to capture the deeper meanings. Challenges still remain for defining the optimal units and interpretations.

3 Semantic Search and Retrieval

Once narratives are embedded, we need semantic retrieval methods that go beyond keyword matching—able to capture story themes, plot arcs, and character dynamics. This is essential for querying things like "reluctant heroes who overcome self-doubt."

3.1 Core Semantic Search & RAG

Semantic search uses vector embeddings to retrieve conceptually similar content. Chunks are embedded and indexed (e.g., FAISS, HNSW) for fast Approximate Nearest Neighbor search [18]. The quality of these embeddings is crucial—poor representation means poor results. Retrieval-Augmented Generation (RAG) [5, 44] pairs this with LLMs. A query is embedded, relevant chunks retrieved, and prepended to the prompt for grounded generation. RAG variants improve on this with better chunking, query rewriting, reranking, and modular retriever-generator setups. **HyDE** [14] improves recall by generating a hypothetical answer, embedding it, and retrieving based on that instead of the query.

3.2 Advanced RAG Variants

Several newer RAG forms handle context more adaptively:

- **FLARE** [19] retrieves mid-generation, only when needed.
- **Self-Routing RAG** [37] skips retrieval if the LLM knows the answer.
- **RetroLM** [29] injects retrieved content into transformer layers directly.

Table 1: Comparison of Selected Long Document Embedding Models and Techniques (Illustrative)

Model/Technique	Max Context	Key Features	Relevant Paper(s)
LongEmbed (E5-NTK)	up to 32k	Training-free context window extension (NTK-aware interpolation for RoPE models)	[47]
StoryEmb	4096 (Mistral base)	Narrative-focused via contrastive learning on summaries; prioritizes plot similarity	[17]
INSTRUCTOR	Model-dependent (e.g., XL up to 512)	Instruction-finetuned; task/domain adaptable via prompts	[34]
HEAL	Model-dependent	Hierarchical alignment via HNMF + contrastive loss for domain-specific retrieval	[1]
HyDE	Encoder-dependent (e.g., Contriever)	Zero-shot dense retrieval via hypothetical document generation	[14]

- **Hierarchical RAG** adds multi-level indexing (e.g., chapters → paragraphs).
- **GraphRAG** [44] brings in structured knowledge via knowledge graphs, enabling graph-enhanced queries like "mentored-by-X and opposed-Y".

3.3 Plot-Based Retrieval

Some models are tailored for plot similarity:

- **FABULA** [33] encodes stories as Event Plot Graphs for structure-aware search.
- **StoryEmb** [17] learns embeddings from plot pairs (e.g., remakes) via contrastive learning.
- **AIStorySimilarity** [4] extracts narrative features (character, theme, events) and compares them via LLM-guided rubrics.

These tools move beyond matching text to retrieving *narrative meaning*, which is critical for long-form fiction search.

4 Advanced Narrative Analysis Techniques

Beyond retrieval, narrative analysis demands identifying themes, tracking characters, and revealing structure. Here we outline scalable methods for deep narrative understanding.

4.1 Extreme Multi-Label Tagging and Few/Zero-Shot Methods

Assigning multiple labels (e.g. #isekai, #antihero, #coming_of_age) is an Extreme Multi-Label Text Classification (XMTC) problem with thousands of sparse, interdependent tags [6]. Major strategies include:

- **Embedding-based**: Project texts and labels into a shared space or compress labels.
- **Hierarchical (Tree-based)**: Cluster instances or labels to prune search paths.
- **Deep Learning**: Transformer models (e.g. AttentionXML) capture dependencies; optimized linear methods remain competitive.
- **LLM-driven**: Quantized models for sampling (QUEST [46]), span-based NER for open labels (GLiNER [43]), or multi-step Infer–Retrieve–Rank pipelines [48].

Building large labeled corpora is costly, motivating zero/few-shot techniques [3]:

- **Direct Prompting** [3]: Simple but biased toward frequent tags.
- **Instruction-Tuned LLMs**: Fine-tuned on diverse tasks, they generalize better (X-Shot [39]).
- **Taxonomy Induction**: LLMs generate label hierarchies (TnT-LLM [36]) for pseudo-labeling.
- **Generative Augmentation** [3]: Synthesizing examples for rare tags.
- **Structured Decomposition**: Break tagging into candidate generation, retrieval, and ranking [48].

4.2 Task Decomposition and Cross-Lingual Analysis

Complex queries (e.g. "Compare protagonist vs. antagonist arcs") benefit from breaking into subtasks: entity extraction, retrieval, sentiment analysis, comparison, and synthesis. Frameworks like AtomR [38] define atomic actions (Search, Filter, Compare, Extract) orchestrated by an LLM controller for interpretable pipelines. Cross-lingual narrative analysis adds challenges beyond translation:

- **Multilingual Models**: mT5, XLM-R, BLOOM handle many languages but vary in performance; benchmarks like MMTEB gauge alignment [10].
- **Cross-Lingual Search**: Query in one language and retrieve across others via aligned embeddings.
- **Machine Translation**: Standard MT often flattens style and misrenders cultural nuance, harming downstream tasks.
- **Knowledge-Enhanced MT**: Use multilingual knowledge graphs for consistent entity translation [5].
- **Cross-Lingual Summaries**: Preserve themes and plots when summarizing in a different language (e.g. CLCTS [45], MTXLS [31]).

Success requires models sensitive to cultural context, narrative structure, and consistent entity grounding across languages.

5 Comparing Methodologies

Choosing between Retrieval-Augmented Generation (RAG), Long Context (LC) models, and Fine-Tuning (FT) hinges on performance needs, cost, data, and adaptability.

5.1 RAG vs. Long Context LLMs

No one-size-fits-all; each excels under different conditions [24, 25].

- **Performance:** LC models shine on dense, coherent texts that fit in their window; RAG outperforms when facts are scattered across many documents or for ultra-long inputs (>64k tokens) where attention degrades [23, 29].
- **Context Handling:** LC attends to everything at once; RAG retrieves and injects only relevant chunks, reducing noise if retrieval is accurate.
- **Failure Modes:** LC may overlook buried details ("needle in haystack") [25]; RAG's bottleneck is retrieval—wrong chunks lead to wrong answers [29].
- **Cost:** LC inference cost grows with input length; RAG pays up-front for indexing but uses smaller contexts per query [32].
- **Model Strength:** Weaker LLMs gain more from RAG's external knowledge; stronger LLMs leverage LC's full context when needed [24].
- **Hybrid Strategies:** Dynamic systems like Self-Route switch between RAG and LC based on query complexity [26].

5.2 Fine-Tuning vs. Inference-Time Augmentation

Decide whether to bake knowledge into model parameters (FT) or supply context at query time (RAG/LC).

Fine-Tuning (FT).

- **Pros:** Embeds domain knowledge or style (e.g. author's voice) into the model; faster inference without large contexts.
- **Cons:** Needs curated data and compute; risk of forgetting general knowledge; updates require retraining.

Inference-Time Augmentation (RAG/LC).

- **Pros:** No retraining for new information; dynamic, cite-able sources; flexible for evolving corpora [5].
- **Cons:** Dependent on retrieval quality or LLM's long-context handling; harder to enforce consistent style; LC remains resource-intensive [32].

Hybrid FT+RAG systems—e.g. RankRAG—fine-tune ranking or style components while using RAG for facts [42]. FT suits style-driven generation; RAG/LC best for content-grounded analysis.

6 Evaluation Strategies

Assessing narrative-understanding systems requires metrics aligned to goals—retrieval accuracy, generation quality, tagging precision, or structural insight—while contending with subjectivity and context-dependence.

6.1 Semantic Search and RAG

RAG splits into retrieval and generation, each with its own metrics:

Retrieval. When relevance labels exist, apply IR metrics:

- **Precision@k**, **Recall@k**, **MRR**, **nDCG@k**, **Hit Rate**
- For chunk-based systems, span overlap (e.g. IoU) gauges retrieval granularity.

Generation. Judge outputs on:

- **Faithfulness**: fidelity to retrieved evidence
- **Relevance**: completeness and focus on the query
- **Correctness**: factual accuracy where verifiable

Automatic metrics (BLEU, ROUGE) correlate poorly for creative text; embedding metrics (BERTScore) improve semantic matching but miss coherence and logic [41]. "LLM-as-Judge" paradigms prompt a strong model to rate faithfulness and relevance [4], though they inherit LLM biases.

6.2 Narrative-Specific Evaluation and Benchmarks

Storytelling demands specialized dimensions beyond n-gram overlap:

- **Coherence & Continuity**: logical flow, consistent character/world details
- **Plot Similarity**: shared event structures and thematic arcs (e.g. AIStorySimilarity [4])
- **Character Arc Tracking**: mapping development trajectories via dynamic embeddings [21]
- **Emotional Progression & Tone**: tracking affective arcs and stylistic consistency
- **Thematic Cohesion**: identifying and maintaining core themes
- **Human Evaluation** [17]: pairwise or Likert ratings remain the standard for engagement, creativity, and plausibility despite cost and annotator variance

Existing benchmarks (MMTEB [10], LONGEMBED [47], LaRA [24], AIStorySimilarity [4], Movie Remake [17], REGEN [35]) cover IR, long-context, or limited narrative tasks, but deep annotations of plot structures, character roles, and thematic elements in large public corpora are scarce—forcing proxy tasks (e.g., from XMTC datasets [6]), human studies, or LLM judgments and hampering standardized comparison.

7 Challenges and Open Problems

Analyzing long-form fiction at scale faces several intertwined barriers:

- **Context vs. Cost:** Holistic understanding demands large context windows or extensive RAG indices, yet both incur high compute and financial expense [32].
- **Plot Modeling:** Current proxies (summaries, chunk embeddings) miss causal, temporal, hierarchical story structure, and struggle with non-linear timelines or unreliable narrators [17, 33].
- **Data Scarcity:** Few public corpora are annotated with detailed narrative schemas (roles, plot points, thematic threads), forcing reliance on zero/few-shot methods [6].
- **Underdeveloped Evaluation:** Standard metrics fail to capture coherence, plot similarity, or thematic depth; robust benchmarks and protocols remain lacking [4, 41].

- **Temporal Fragility:** Modeling long-range character evolution, shifting relationships, and sudden plot pivots—especially across multiple timelines or flashbacks—exceeds current temporal reasoning capabilities [18, 21].
- **Scaling Advanced Methods:** Graph-based RAG and complex XMTC pipelines show promise but face engineering hurdles indexing and querying millions of narratives efficiently [46].
- **Explainability:** Beyond attention weights, narrative systems need XAI techniques that clarify why particular stories are retrieved, tags assigned, or summaries generated [16].
- **Bias & Fairness:** LLMs can amplify stereotypes or marginalize voices; detecting and mitigating such biases is crucial for equitable narrative analysis [13, 28, 40, 49].
- **Humanistic Insight Gap:** Bridging pattern-based analysis with deep literary interpretation (symbolism, irony, intertextuality) demands interdisciplinary collaboration.

8 Conclusion

Semantic search and analysis for long-form narrative fiction is a complex systems challenge: it demands representations that capture plot structure under tight context limits, scalable retrieval and generation across multilingual corpora, and architectures integrating embeddings, RAG, task decomposition, and literary insight. We surveyed length-aware embeddings (LongEmbed [47], StoryEmb [17], INSTRUCTOR [11, 34]), temporal modeling techniques [18, 21], advanced RAG pipelines (FLARE [19], GraphRAG [44]), XMTC and zero/few-shot tagging [3, 6], and decomposition frameworks [38]. Despite progress, plot modeling remains rudimentary, evaluation metrics underrepresent narrative depth, and context–cost and bias–fairness trade-offs endure. Future work should blend long-context models for coherence, RAG for grounding, knowledge graphs for relational reasoning, dynamic temporal pipelines, and explainable interfaces to enable scalable narrative scholarship and richer reader engagement.

References

[1] Manish Bhattarai, Ryan Barron, Maksim Eren, Minh Vu, Vesselin Grantcharov, Ismael Boureima, Valentin Stanev, Cynthia Matuszek, Vladimir Valtchinov, Kim Rasmussen, and Boian Alexandrov. 2024. HEAL: Hierarchical Embedding Alignment Loss for Improved Retrieval and Representation Learning. arXiv:2412.04661 [cs.IR] https://arxiv.org/abs/2412.04661

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[3] Georgios Chochlakis, Efthymios Georgiou, and Alexandros Potamianos. 2021. End-to-end Generative Zero-shot Learning via Few-shot Learning. arXiv:2102.04379 [cs.CV] https://arxiv.org/abs/2102.04379

[4] Jon Chun. 2024. AIStorySimilarity: Quantifying Story Similarity Using Narrative for Search, IP Infringement, and Guided Creativity. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, Libby Barak and Malihe Alikhani (Eds.). Association for Computational Linguistics, Miami, FL, USA, 161–177. doi:10.18653/v1/2024.conll-1.13

[5] Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards Cross-Cultural Machine Translation with Retrieval-Augmented Generation from Multilingual Knowledge Graphs. arXiv:2410.14057 [cs.CL] https://arxiv.org/abs/2410.14057

[6] Arpan Dasgupta, Siddhant Katyan, Shrutimoy Das, and Pawan Kumar. 2023. Review of Extreme Multilabel Classification. arXiv:2302.05971 [cs.LG] https://arxiv.org/abs/2302.05971

[7] Xingjian Diao, Chunhui Zhang, Weiyi Wu, Zhongyu Ouyang, Peijun Qing, Ming Cheng, Soroush Vosoughi, and Jiang Gui. 2025. Temporal Working Memory: Query-Guided Segment Refinement for Enhanced Multimodal Understanding. arXiv:2502.06020 [cs.CV] https://arxiv.org/abs/2502.06020

[8] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130 [cs.CL] https://arxiv.org/abs/2404.16130

[9] Jan Elfes. 2024. Mapping News Narratives Using LLMs and Narrative-Structured Text Embeddings. arXiv:2409.06540 [cs.CL] https://arxiv.org/abs/2409.06540

[10] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark. arXiv:2502.13595 [cs.CL] https://arxiv.org/abs/2502.13595

[11] Hongjin Su et al. 2024. INSTRUCTOR. https://instructor-embedding.github.io/. Accessed: 2025-04-20.

[12] Yichao Feng, Shuai Zhao, Yueqiu Li, Luwei Xiao, Xiaobao Wu, and Anh Tuan Luu. 2025. Aspect-Based Summarization with Self-Aspect Retrieval Enhanced Generation. arXiv:2504.13054 [cs.CL] https://arxiv.org/abs/2504.13054

[13] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. arXiv:2309.00770 [cs.CL] https://arxiv.org/abs/2309.00770

[14] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise Zero-Shot Dense Retrieval without Relevance Labels. arXiv:2212.10496 [cs.IR] https://arxiv.org/abs/2212.10496

[15] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL] https://arxiv.org/abs/2312.10997

[16] Mareike Hartmann, Han Du, Nils Feldhus, Ivana Kruijff-Korbayová, and Daniel Sonntag. 2022. XAINES: Explaining AI with Narratives. *KI - Künstliche Intelligenz* 36, 3 (01 Dec 2022), 287–296. doi:10.1007/s13218-022-00780-8

[17] Hans Ole Hatzel and Chris Biemann. 2024. Story Embeddings — Narrative-Focused Representations of Fictional Stories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 5931–5943. doi:10.18653/v1/2024.emnlp-main.339

[18] Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Dynamic Contextualized Word Embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6970–6984. doi:10.18653/v1/2021.acl-long.542

[19] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. arXiv:2305.06983 [cs.CL] https://arxiv.org/abs/2305.06983

[20] A. Joshi, E. Fidalgo, E. Alegre, and R. Alaiz-Rodriguez. 2024. RankSum An unsupervised extractive text summarization based on rank fusion. arXiv:2402.05976 [cs.LG] https://arxiv.org/abs/2402.05976

[21] Vani K, Simone Mellace, and Alessandro Antonucci. 2020. Temporal Embeddings and Transformer Models for Narrative Text Understanding. arXiv:2003.08811 [cs.CL] https://arxiv.org/abs/2003.08811

[22] Srijan Kumar, Xikun Zhang, and Jure Leskovec. 2019. Predicting Dynamic Embedding Trajectory in Temporal Interaction Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. ACM, 1269–1278. doi:10.1145/3292500.3330895

[23] Quinn Leng, Jacob Portes, Sam Havens, Matei Zaharia, and Michael Carbin. 2024. Long Context RAG Performance of Large Language Models. arXiv:2411.03538 [cs.LG] https://arxiv.org/abs/2411.03538

[24] Kuan Li, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang, and Minhao Cheng. 2025. LaRA: Benchmarking Retrieval-Augmented Generation and Long-Context LLMs – No Silver Bullet for LC or RAG Routing. arXiv:2502.09977 [cs.CL] https://arxiv.org/abs/2502.09977

[25] Xinze Li, Yixin Cao, Yubo Ma, and Aixin Sun. 2024. Long Context vs. RAG for LLMs: An Evaluation and Revisits. arXiv:2501.01880 [cs.CL] https://arxiv.org/abs/2501.01880

[26] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. arXiv:2407.16833 [cs.CL] https://arxiv.org/abs/2407.16833

[27] Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, and Jingrui He. 2025. Language in the Flow of Time: Time-Series-Paired Texts Weaved into a Unified Temporal Narrative. arXiv:2502.08942 [cs.LG] https://arxiv.org/abs/2502.08942

[28] Zhaoming Liu. 2024. Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies. *Journal of Transcultural Communication* (2024). doi:doi:10.1515/jtc-2023-0019

[29] Kun Luo, Zheng Liu, Peitian Zhang, Hongjin Qian, Jun Zhao, and Kang Liu. 2025. Does RAG Really Perform Bad For Long-Context Processing? arXiv:2502.11444 [cs.CL] https://arxiv.org/abs/2502.11444

[30] Asmar Nadeem, Faegheh Sardari, Robert Dawes, Syed Sameed Husain, Adrian Hilton, and Armin Mustafa. 2025. NarrativeBridge: Enhancing Video Captioning with Causal-Temporal Narrative. arXiv:2406.06499 [cs.CV] https://arxiv.org/abs/2406.06499

[31] Diogo Pernes, Gonçalo M. Correia, and Afonso Mendes. 2024. Multi-Target Cross-Lingual Summarization: a novel task and a language-neutral approach. arXiv:2410.00502 [cs.CL] https://arxiv.org/abs/2410.00502

[32] Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers. arXiv:2404.04925 [cs.CL] https://arxiv.org/abs/2404.04925

[33] Priyanka Ranade and Anupam Joshi. 2023. FABULA: Intelligence Report Generation Using Retrieval-Augmented Narrative Construction. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM '23)*. ACM, 603–610. doi:10.1145/3625007.3627505

[34] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. arXiv:2212.09741 [cs.CL] https://arxiv.org/abs/2212.09741

[35] Kun Su, Krishna Sayana, Hubert Pham, James Pine, Yuri Vasilevski, Raghavendra Vasudeva, Marialena Kyriakidi, Liam Hebert, Ambarish Jash, Anushya Subbiah, and Sukhdeep Sodhi. 2025. REGEN: A Dataset and Benchmarks with Natural Language Critiques and Narratives. arXiv:2503.11924 [cs.CL] https://arxiv.org/abs/2503.11924

[36] Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W White, Longqi Yang, Reid Andersen, Georg Buscher, Dhruv Joshi, and Nagu Rangan. 2024. TnT-LLM: Text Mining at Scale with Large Language Models. arXiv:2403.12173 [cs.CL] https://arxiv.org/

abs/2403.12173
[37] Di Wu, Jia-Chen Gu, Kai-Wei Chang, and Nanyun Peng. 2025. Self-Routing RAG: Binding Selective Retrieval with Knowledge Verbalization. arXiv:2504.01018 [cs.CL] https://arxiv.org/abs/2504.01018
[38] Amy Xin, Jinxin Liu, Zijun Yao, Zhicheng Lee, Shulin Cao, Lei Hou, and Juanzi Li. 2025. AtomR: Atomic Operator-Empowered Large Language Models for Heterogeneous Knowledge Reasoning. arXiv:2411.16495 [cs.CL] https://arxiv.org/abs/2411.16495
[39] Hanzi Xu, Muhao Chen, Lifu Huang, Slobodan Vucetic, and Wenpeng Yin. 2024. X-Shot: A Unified System to Handle Frequent, Few-shot and Zero-shot Learning Simultaneously in Classification. arXiv:2403.03863 [cs.CL] https://arxiv.org/abs/2403.03863
[40] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* 55, 1 (2024), 90–112. doi:10.1111/bjet.13370 arXiv:https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13370
[41] Dingyi Yang and Qin Jin. 2024. What Makes a Good Story and How Can We Measure It? A Comprehensive Survey of Story Evaluation. arXiv:2408.14622 [cs.CL] https://arxiv.org/abs/2408.14622
[42] Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. arXiv:2407.02485 [cs.CL] https://arxiv.org/abs/2407.02485
[43] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. arXiv:2311.08526 [cs.CL] https://arxiv.org/abs/2311.08526
[44] Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Junnan Dong, Hao Chen, Yi Chang, and Xiao Huang. 2025. A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models. arXiv:2501.13958 [cs.CL] https://arxiv.org/abs/2501.13958
[45] Ran Zhang, Jihed Ouni, and Steffen Eger. 2024. Cross-lingual Cross-temporal Summarization: Dataset, Models, Evaluation. arXiv:2306.12916 [cs.CL] https://arxiv.org/abs/2306.12916
[46] Chuang Zhou, Junnan Dong, Xiao Huang, Zirui Liu, Kaixiong Zhou, and Zhaozhuo Xu. 2024. QUEST: Efficient Extreme Multi-Label Text Classification with Large Language Models on Commodity Hardware. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 3929–3940. doi:10.18653/v1/2024.findings-emnlp.226
[47] Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. LongEmbed: Extending Embedding Models for Long Context Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 802–816. doi:10.18653/v1/2024.emnlp-main.47
[48] Yaxin Zhu and Hamed Zamani. 2024. ICXML: An In-Context Learning Framework for Zero-Shot Extreme Multi-Label Classification. arXiv:2311.09649 [cs.LG] https://arxiv.org/abs/2311.09649
[49] Li Zhui, Li Fenghe, Wang Xuehu, Fu Qining, and Ren Wei. 2024. Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint. *J Med Internet Res* 26 (1 Aug 2024), e60083. doi:10.2196/60083