Data Foundations of Long-Context Language Models: A Survey

Zechen Sun[1], Yuyang Sun[1], Zhaochen Su[2], Zecheng Tang[1], Juntao Li[1]*, Ao Zhou[3], Wenliang Chen[1], Min Zhang[1]

[1]Institute of Computer Science and Technology, Soochow University
[2]Hong Kong University of Science and Technology
[3]State Key Laboratory of Novel Software Technology, Nanjing University
{zcsunc, yysuns, zctang}@stu.suda.edu.cn, suzhaochen0110@gmail.com
{ljt, wlchen, minzhang}@suda.edu.cn, za@smail.nju.edu.cn

Abstract

As the context window of Large Language Models (LLMs) continues to expand, the data required to effectively train and evaluate these capabilities remains underexplored. With existing research primarily focuses on architectural optimization, there is a need for a systematic, data-centric review. This survey bridges this gap by investigating the data foundations of Long-Context Language Models (LCMs). We begin by examining current data strategies alongside their strengths and limitations, mapping the required data to desired model capabilities. Building on this, we explore how targeted training data designs drive core, often interconnected skills such as retrieval, reasoning, and aggregation. Furthermore, we analyze the evaluation landscape, illustrating how selecting appropriate benchmarks is crucial for probing capability boundaries and guiding effective model selection. Finally, we synthesize actionable guidelines for data construction and outline critical future directions to propel the advancement of long-context language models, including quantifying data quality, establishing scaling laws for length distributions, and developing dynamic evaluation frameworks.

1 Introduction

Long-Context Language Models (LCMs), typically defined as models with a context window exceeding 128K tokens, have gained significant attention in natural language processing (Comanici et al., 2025; Team et al., 2026; GLM-5-Team, 2026). However, effective long-context modeling requires more than simply supporting an extended input context window. In practice, LCMs are expected to process long sequences comprehensively and accurately. This involves tasks requiring a global understanding of the text, such as precisely retrieving relevant information or aggregating dispersed content across long documents (Kim et al., 2024; Roberts et al., 2025; Xiao et al., 2026).

Developing models to handle such extensive sequences presents significant computational challenges (Shi et al., 2025; Li et al., 2026; Shen and Shen, 2026). Consequently, existing surveys primarily focus on techniques for extending context windows (Wang et al., 2024c) and efficient LCM training (Liu et al., 2025b). Despite this progress, a clear gap remains: a comprehensive review focused on long-context data strategies is missing. The capabilities of LCMs are fundamentally driven by their training data. Compared to traditional short-text corpora, long-context data has unique characteristics, such as extended structural dependencies, non-uniform information density, and complex length distributions (Deng et al., 2025; Wang et al., 2025b; Alla et al., 2026). These factors require specialized data construction and utilization strategies that have not been fully reviewed.

To bridge this gap and better understand how data influences model performance, this survey examines the relationship between long-context data and specific model capabilities. We categorize the core capabilities of LCMs into five areas: language modeling, retrieval, reasoning, aggregation, and long-form generation. Although these capabilities overlap in practice (Li et al., 2024b; Bertsch et al., 2025), outlining them helps illustrate how datasets target different skills. Specifically, ***different training data designs provide distinct gains across these capabilities, and evaluation benchmarks are built with corresponding focuses***. Selecting the appropriate evaluation data is essential to probe capability boundaries and guide model selection (Zhang et al., 2025a; Chen et al., 2026b; Xiao et al., 2025; Zeng et al., 2026).

* Juntao Li is the Corresponding Author.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1803–1825, 2026. https://doi.org/10.1162/TACL.a.775
Action Editor: Pradeep Dasigi. Submission batch: 11/2025; Revision batch: 3/2026; Published 7/2026.

Structure of the Survey

This survey systematically organizes and analyzes the data foundations of LCMs. The overall taxonomy of long-context data is visualized in Figure 1. Table 1 and Table 2 summarize existing training datasets and evaluation benchmarks across five dimensions: Source, Stage, Construction Methods, Length Distribution and Capabilities. The survey is organized as follows:

- **Section 2** examines the required characteristics of long-context data and maps them to the expected capabilities for effective LCM training and evaluation.
- **Section 3** categorizes the sources of long-context data. This section details their respective construction methods, and discusses their strengths and limitations.
- **Section 4** details how specific training data designs drive core long-context skills and examines how targeted benchmarks evaluate these corresponding capabilities.
- **Section 5** outlines key data-centric challenges and proposes future directions, including quantifying data quality, analyzing length distributions, and developing dynamic evaluation frameworks.

2 Requirements for Long-Context Data

Beyond merely accepting extended inputs (such as 128K tokens), long-context language models (LCMs) must comprehensively understand the entire provided text (Deng et al., 2025). This requires the ability to precisely retrieve key information and to reason over or aggregate widely dispersed points within extremely long contexts. Furthermore, LCMs are expected to generate sufficiently long and globally coherent outputs, such as novels or reports exceeding 32K tokens. Therefore, the design of both training and evaluation data should explicitly target these expected capabilities.

2.1 Training

Building upon these requirements, the construction of training data must satisfy several key properties. First, training data must maintain strong semantic coherence to avoid fragmented contexts. Incoherent data can cause models to incorrectly link unrelated facts, which increases the risk of reasoning mistakes (Liu et al., 2024d; Chen et al., 2024b). Second, the corpus needs abundant long-range dependencies. This helps the model capture semantic relationships across distant paragraphs and resolve complex coreferences. Consequently, it prevents the model from forgetting earlier information and mitigates the well-known "lost in the middle" phenomenon (Chen et al., 2024b; Baker et al., 2024). Finally, data diversity is crucial for improving model generalization. The training corpus must cover multiple domains (e.g., science, literature, and news) and genres (e.g., academic papers, novels, and reports). It should also include structured formats such as code and tables to enhance the ability of the model to understand complex long documents (Fu et al., 2024; Liu et al., 2024b; Ye et al., 2025a; Staniszewski et al., 2025).

2.2 Evaluation

Evaluation benchmarks must specifically target the core capabilities expected of LCMs. A well-designed evaluation helps explore model capability boundaries, reflects real-world performance, and drives further progress in long-context processing. Specifically, these benchmarks should cover the following key dimensions. First, Information Retrieval Accuracy tests whether the model can accurately recall specific facts from massive contexts (Kuratov et al., 2024; Xu et al., 2024). Needle-in-a-Haystack (NIAH) tests extend this by measuring retrieval robustness as the input length continuously grows. Second, Multi-hop Reasoning in long-contexts examine the ability to integrate distributed information and perform complex logic inference (Zhu et al., 2024; Zhang et al., 2025c). Third, Long-context Summarization and Information Aggregation tasks evaluate how well the model synthesizes global content and extracts core concepts from lengthy input sequences (Yuan et al., 2024). Finally, Long-form Generation tasks assess the capacity of the model to produce long-form and high-quality outputs. Evaluating such lengthy generated content requires a comprehensive set of metrics that includes relevance, accuracy, coherence, clarity, breadth and depth (Bai et al., 2025b; Wu et al., 2025b; Wang et al., 2025a).

3 Obtaining Long-Context Data

As discussed in Section 2, LCMs require data that features abundant long-range dependencies and strong semantic coherence. Acquiring such data is challenging. To address this, current research generally relies on two approaches. The first approach involves sourcing naturally occurring long texts

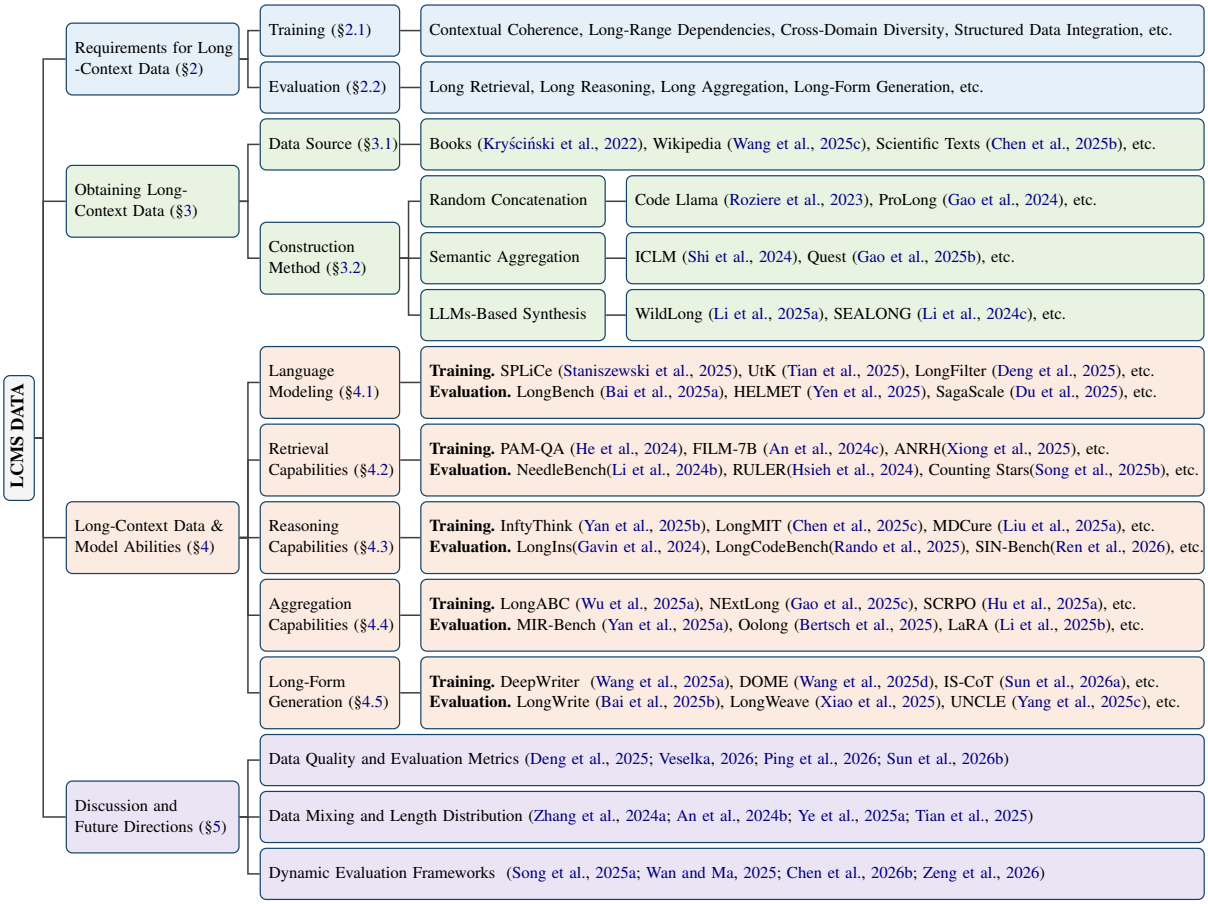

Figure 1: Taxonomy of Long-Context Data. It provides a structured overview of the field, encompassing data requirements and sourcing methods, with a particular focus on the mapping between specific data strategies and desired long-context capabilities.

that inherently possess contextual relationships, which we detail in §3.1. The second approach recognizes that raw natural data often cannot be used directly, and manual annotation for extremely long texts is costly. Therefore, researchers apply specific construction strategies and advanced models to synthesize high-quality training and evaluation data, which we discuss in §3.2.

3.1 Natural Long-Context Data Source

General Texts. General texts are fundamental resources due to their large scale, high diversity, and ease of acquisition (Zhao et al., 2023). These characteristics make them highly suitable for improving the basic language modeling and generalization capabilities of LCMs. **Books** serve as a primary natural source of high-quality long-context data (Kim et al., 2024). Their diverse genres, coherent narratives, and complex textual relationships, make them highly effective for enhancing language understanding (Bai et al., 2024a). Widely used corpora include Books3 and BookCorpus2[1]. For example, LADM (Chen et al., 2025b) selects long documents from book corpora to support LCM pretraining. **Wikipedia**[2], as a structured knowledge resource with broad topic coverage, is also widely used. For instance, LongPAS (Peng et al., 2026) constructs knowledge graphs from Wikipedia and generates multi-hop reasoning chains to enhance the reasoning capabilities of LCMs.

Domain-specific Data. Domain-specific data is essential to optimize and evaluate model performance on specialized long-context tasks (Yan et al., 2025b; Adams et al., 2025). **Scientific literature** is a core category of domain-specific data, encompassing academic papers and textbooks. Compared to general books, scientific literature is crucial for scientific knowledge understanding and specialized reasoning capabilities of LCMs (Tian et al., 2025; Chen et al., 2025b). Common sources

[1] https://opendatalab.com/OpenDataLab/Pile-BookCorpus2

[2] https://zh.wikipedia.org/

Datasets	Source	Stage	Construction Methods	Length
Language Modeling				
Quest (Gao et al., 2025b)	O	Pre-training	Splicing	$\leq$ 1M
SPLiCe (Staniszewski et al., 2025)	W, S, C	Pre-training	Splicing	$\leq$ 32K
UtK (Tian et al., 2025)	B, W, S, C	Pre-training	Splicing	$\leq$ 128K
MMR&FPS (Wang et al., 2025e)	O	Pre-training	Generation	-
LongFilter (Deng et al., 2025)	O	Pre-training	Generation	8K~64K
QwenLong-L1.5 (Shen et al., 2025)	B, W, S, C, O	Pre-training	Generation	~256K
LiteLong (Jia et al., 2025)	O	Pre-training	Generation, Splicing	~128K
LADM (Chen et al., 2025b)	B, W, S, C, O	Pre-training	Generation, Sampling	~32K
NextLong (Gao et al., 2025c)	B, W, S	Pre-training	Posi-Synthesis, Splicing	$\leq$ 128K
LongABC (Wu et al., 2025a)	B, C, S	Pre-training	Posi-Synthesis, Sampling	~32K
Retrieval				
MMLBD-C (Veselka, 2026)	S, O	Post-training	Annotation	~344K
Arti-Needles (Xiong et al., 2025)	O	Post-training	Generation	$\leq$ 4K
LongPAS (Peng et al., 2026)	W	Post-training	Generation	~60K
MegaBeam (Wu and Song, 2025)	B, S, C	Post-training	Splicing, Posi-Synthesis	$\leq$ 512K
Reasoning				
ReconstructionLong (Xiao et al., 2026)	B, W, C	Post-training	Annotation	~64K
LongMIT (Chen et al., 2025c)	B, W, S	Post-training	Generation	4K~128K
USDC (Marreddy et al., 2025)	O	Post-training	Generation	$\leq$ 4K
LongFaith (Yang et al., 2025b)	W	Post-training	Generation	~9K
InftyThink (Yan et al., 2025b)	O	Post-training	Generation	-
Light-R1 (Wen et al., 2025)	O	Post-training	Generation	$\leq$ 32K
SSB (Mitra and Ulukus, 2025)	O	Post-training	Generation	-
LongPLVR (Chen et al., 2026a)	O	Post-training	Generation	8K~64K
LongR (Ping et al., 2026)	C, O	Post-training	Generation, Mixing	16K~32K
LongMagpie (Gao et al., 2025a)	O	Post-training	Generation, Splicing	~64K
HIERARCHICAL (He et al., 2025)	B	Post-training	Generation, Posi-Synthesis	$\leq$ 1M
Aggregation				
MDCure (Liu et al., 2025a)	O	Post-training	Generation	$\leq$ 32K
WildLong (Li et al., 2025a)	S, C	Post-training	Generation	-
LongReward (Zhang et al., 2025b)	S	Post-training	Generation	8K~64K
SCRPO (Hu et al., 2025a)	O	Post-training	Generation	-
LOGO (Tang et al., 2025)	B, W, S	Post-training	Generation, Posi-Synthesis	$\leq$ 80K
Long-form Generation				
OpenReward (Hu et al., 2025b)	W, S, O	Post-training	Mixing	~6K
PrefBERT (Li et al., 2025d)	O	Post-training	Mixing, Generation	~4K
DOME (Wang et al., 2025d)	B, O	Post-training	Generation	-
LongWriter (Bai et al., 2025b)	O	Post-training	Generation	2K~32K
LongDPO (Ping et al., 2025)	O	Post-training	Generation	$\leq$ 32K
DeFine (Wang et al., 2025c)	W	Post-training	Generation	-
DeepWriting (Wang et al., 2025a)	O	Post-training	Generation	-
Writing-RL (Lei et al., 2025)	O	Post-training	Generation	$\leq$ 8K
SuperWriter (Wu et al., 2025b)	O	Post-training	Generation	-
ACE-RL (Chen et al., 2025a)	C, O	Post-training	Generation	~8K
ProxyReward (Guo et al., 2025)	S, O	Post-training	Generation	~2K

Table 1: Datasets for LCMs Training categorized by parallel target capabilities. "B", "W", "S", "C", and "O" refer to Books, Wikipedia, Scientific-Texts, Code, and Other domain-specific data. The "Length" column indicates the token length distribution. Due to the interdependence of long-context capabilities, many datasets concurrently target multiple overlapping skills. To align with the structure of Section 4, each dataset is placed in the block corresponding to its most prominent capability.

include arXiv[3] and PubMed[4]. **Code** is another critical category because it contains precise logical structures and strict long-range dependencies. Primary sources include programming question-answer communities like Stack Exchange[5], which provide practical application scenarios, and public software repositories like GitHub[6], which offer diverse programming languages and large-scale projects (Wu and Song, 2025; Wu et al., 2025a).

Acquiring these general and domain-specific corpora is the first step in the data pipeline; processing them into training and evaluation formats needs additional construction methods.

[3] https://arxiv.org/
[4] https://pubmed.ncbi.nlm.nih.gov/
[5] https://stackexchange.com/
[6] https://github.com/

Benchmarks	Source	Construction Method	Length
Multi-task Benchmarks			
Ruler (Hsieh et al., 2024)	O	Generation	Configurable
U-NIAH (Gao et al., 2025d)	O	Generation	-
SagaScale (Du et al., 2025)	B, W	Generation	~228K
L-EVAL (An et al., 2024a)	B, O	Annotation	3K~200K
HELMET (Yen et al., 2025)	B, W, O	Annotation	$\leq$ 128K
LongBench Pro (Chen et al., 2026b)	B, S, C, O	Annotation	8K~32K
LV-EVAL (Yuan et al., 2024)	B, W, O	Annotation, Generation	16K~256K
LongBench v1&v2 (Bai et al., 2024b, 2025a)	B, W, S	Annotation, Generation	8K~2M
InftyBench (Zhang et al., 2024b)	B, C	Annotation, Generation	~200K
LongSafety (Lu et al., 2025)	S, O	Annotation, Generation	~5K
Retrieval Benchmarks			
Needle Threading (Roberts et al., 2025)	B	Generation	$\leq$ 630K
LOCA-Bench (Zeng et al., 2026)	B, O	Generation	16K~256K
Task Haystack (Xu et al., 2024)	W, O	Annotation	4K~32K
StoryBench (Wan and Ma, 2025)	O	Annotation	-
BABILong (Kuratov et al., 2024)	B	Annotation, Generation	$\leq$ 10M
NoLiMa (Modarressi et al., 2025a)	B	Annotation, Generation	$\leq$ 2M
Reasoning Benchmarks			
XL2Bench (Ni et al., 2024)	B, S, O	Generation	100K~200K
LongIns (Gavin et al., 2024)	O	Generation	$\leq$ 16K
HoloBench (Maekawa et al., 2025)	W, O	Generation	$\leq$ 128K
LongReasonArena (Ding et al., 2025)	O	Generation	$\leq$ 1M
LongReason (Ling et al., 2025)	B, O	Generation	Configurable
MemoryRewardBench (Tang et al., 2026)	O	Generation	8K~128K
FanOutQA (Zhu et al., 2024)	W	Annotation	-
FinTextQA (Chen et al., 2024a)	O	Annotation	~19.7K
NovelQA (Wang et al., 2024a)	B	Annotation	~200K
Counting-Star (Song et al., 2025c)	O	Annotation	$\leq$ 1M
CURIE (Cui et al., 2025)	S	Annotation	$\leq$ 64K
LaRA (Li et al., 2025b)	B, S, O	Annotation, Generation	32K~128K
LongCodeBench (Rando et al., 2025)	C	Annotation, Generation	$\leq$ 1M
SIN-Bench (Ren et al., 2026)	S	Annotation, Generation	32K~1M
Aggregation Benchmarks			
NarrativeQA (Bohnet et al., 2024)	B	Generation	-
Michelangelo (Vodrahalli et al., 2024)	W, C	Generation	$\leq$ 1M
SummHay (Laban et al., 2024)	O	Generation	~93K
MIR-Bench (Yan et al., 2025a)	C	Generation	-
Loong (Wang et al., 2024b)	S, O	Annotation	10K~200K
LCFO (Costa-jussà et al., 2025)	W, S, O	Annotation	~5K
Oolong (Bertsch et al., 2025)	W, O	Annotation, Generation	~5K
Long-Form Generation Benchmarks			
HelloBench (Que et al., 2024)	B, W, S, O	Annotation	$\leq$ 16K
LongProc (Ye et al., 2025b)	W, C, O	Annotation	0.5K~8K
ExpertLongBench (Ruan et al., 2025)	B, O	Annotation	~5K
LongLaMP (Kumar et al., 2024)	S, O	Generation	$\leq$ 4K
LongInOutBench (Zhang et al., 2025c)	S	Generation	4K~8K
AcademicEval (Zhang et al., 2025a)	S	Generation	Configurable
Instance-Level (Hsu et al., 2025)	C, O	Generation	-
LongWeave (Xiao et al., 2025)	C, O	Generation	~64K
LongGenBench (Liu et al., 2024c)	O	Generation	4K~128K
LongFact (Wei et al., 2024)	O	Generation	-

Table 2: Benchmarks for LCMs Evaluation, grouped by their primary targeted capabilities. "B", "W", "S", "C", and "O" refer to Books, Wikipedia, Scientific-Texts, Code, and Other domain-specific data. The "Length" column details the context window range evaluated.

Summary.

These data sources inherently contain richer long-range dependencies. They are typically combined with specific construction strategies to develop and benchmark LCMs.

3.2 Data Construction Strategy

While natural long texts provide a solid foundation, raw data can rarely be used directly to train or evaluate LCMs. Instead, these texts typically require appropriate reorganization or formatting to serve as model inputs. Moreover, manually annotating in-

structions or designing evaluation queries for such lengthy documents is challenging. As a result, automated construction and synthesis strategies have become the standard practice for acquiring.

Random Concatenation. This method rapidly generates long-context data by randomly combining multiple documents (Ouyang et al., 2022; Le Scao et al., 2023; Touvron et al., 2023). This approach is operationally simple and scales easily, enabling the acquisition of massive text sequences at a low cost. However, because there are no inherent semantic connections between the combined documents, this method disrupts logical structure and textual coherence. Consequently, models primarily learn to adapt to longer input lengths rather than capturing long-range dependencies. Therefore, this method is mostly suitable for the unsupervised pre-training, specifically when researchers need to rapidly extend the model's context window length.

Semantic Aggregation. This method constructs long contexts by grouping semantically similar documents, typically by concatenating a document with its top-k most similar neighbors in the embedding space (Shi et al., 2024; Yang et al., 2024a). This approach maintains a degree of local textual coherence, as adjacent documents share similar topics. However, this increased coherence often leads to information redundancy. Semantically similar documents often contain overlapping content, which reduces the effective information density and limits the diversity of the context. As a result, semantic aggregation is also primarily used in unsupervised pre-training, serving as a compromise between preserving local coherence and maintaining construction efficiency (Gao et al., 2025b; Comanici et al., 2025; Zeng et al., 2025).

LLM-Based Synthesis. For supervised fine-tuning and complex evaluation, simple concatenation is insufficient. LLM-based synthesis leverages advanced models to construct supervised data, effectively solving the high cost of manual annotation (An et al., 2024c; Zhang et al., 2025b; Ping et al., 2025; Sun et al., 2026b). This approach can flexibly generate instruction-tuning and preference-alignment data, often formatted as question-answer pairs designed for long contexts. The main advantages of synthetic data are its high scalability and the flexibility to support various downstream tasks. In scenarios where supervised signals are scarce, LLM synthesis has become the core methodology for acquiring post-training data (Hong et al., 2024; Yang et al., 2024b). However, researchers must also manage potential risks associated with this method, such as model biases and deviations from real-world data distributions.

Summary.

Data construction strategies bridge the gap between raw natural texts and the specific formatting required for LCM. While concatenation methods efficiently scale data for pre-training, LLM-based synthesis has become the standard approach for generating high-quality supervised data.

4 Long-Context Data Driven Model Capabilities

Figure 2 introduces a taxonomy of LCM capabilities, categorized into five parallel dimensions: Language Modeling (§4.1), Retrieval (§4.2), Reasoning (§4.3), Aggregation (§4.4), and Long-Form Generation (§4.5). We claim these capabilities are highly interdependent but have distinct focuses. This taxonomy highlights a central data-centric principle: different training data strategies provide different gains for these specific focuses, and benchmarks are designed to evaluate these different dimensions. By mapping specific data engineering strategies and benchmark designs to each capability, this section provides a clear guide for selecting appropriate datasets to effectively enhance model abilities and perform robust evaluation.

4.1 Language Modeling Capability

The language modeling capability refers to the core understanding abilities essential for processing extended sequences. It emphasizes long-range semantic coherence and information consistency, allowing LCMs to maintain logical integrity throughout long contexts (MiniMax et al., 2025).

These capabilities are primarily developed during pre-training, where data construction and selection strategies play a critical role. Recent approaches generally utilize three data strategies:

- **Splicing and Upsampling.** This approach involves concatenating short natural texts into long sequences and increasing the proportion of high-quality long-context samples (Yang et al., 2025a; GLM-5-Team, 2026). While this method successfully retains natural linguistic structures, it often suffers from a low density of long-range

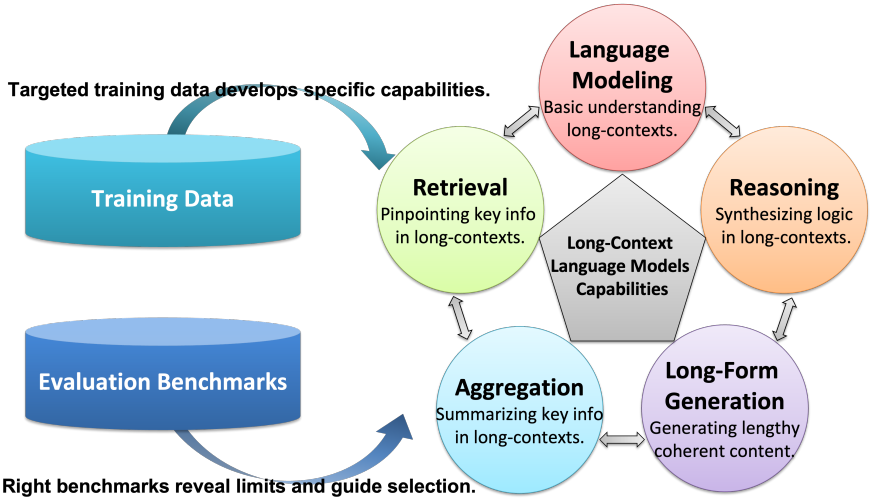

Figure 2: Taxonomy of primary LCMs capabilities and data-driven dependencies. Developing skills requires targeted training corpora, whereas selecting appropriate benchmarks is essential for revealing model limits

dependencies. To mitigate this, filtering frameworks like LongFilter (Deng et al., 2025) evaluate the information gain between extended and truncated contexts. This retains data with abundant long-range dependencies without compromising natural text structures (Zhuang et al., 2025).

- **Segmentation and Reassembly.** To directly increase the dependency density, methods like UtK (Tian et al., 2025) and SPLiCe (Staniszewski et al., 2025) split long documents and reorder the segments. This creates challenging training data that forces the model to attend to distant tokens. However, this strategy may deviate from natural language distributions (Gao et al., 2025b).
- **Domain Mixing Optimization.** Maintaining a balance across different data sources is important for robust generalization (Ye et al., 2025a; Fu et al., 2024). This principle guides the creation of well-proportioned and highly diverse training datasets like LongWanJuan (Liu et al., 2024d) and RegMIX (Liu et al., 2024b).

The evaluation of language modeling is shifting from simple length testing toward analyzing effective context utilization and realistic complexity.

- **Ultra-Long Context Limits.** Benchmarks like RULER (Hsieh et al., 2024) and InftyBench (Zhang et al., 2024b) evaluate context processing exceeding 100K tokens. Building on this, LongBench-Pro (Chen et al., 2026b) reveals a critical insight: the actual effective context length of current models is often significantly shorter than their claimed capacity.
- **Realistic Scenario Complexity.** To evaluate LCMs in practical scenarios, SagaScale (Du et al., 2025) automatically generates complex queries using novels and external knowledge. Furthermore, LongBench (Bai et al., 2025a) and HELMET (Yen et al., 2025) construct datasets based on diverse, application-driven categories.
- **Safety Evaluation.** As expanding context windows introduce new risks, LongSafety (Lu et al., 2025) demonstrates that models become more vulnerable to attacks as input length increases.

Summary.

- **Training.** Current practices suggest a trade-off between coherence (via splicing) and explicitly enforcing long-range dependencies (via reassembly).
- **Evaluation.** Recent benchmarks increasingly focus on measuring effective context utilization, practical scenario complexity, and safety risks in long contexts.

4.2 Retrieval Capability

The retrieval capability refers to the ability to accurately locate and extract key information from long contexts (Roberts et al., 2025). This capability is not simply keyword matching. It requires models to process structured data, filter out noise, and pinpoint semantically relevant segments across massive token spans (Goldman et al., 2024).

A persistent challenge in long-context retrieval is the "lost in the middle" phenomenon, where models fail to access critical content located in the middle of sequences (Liu et al., 2024a; Zhang et al.,

2025c). Recent studies focus on mitigating this during post-training stage via targeted data construction (Xu et al., 2025; Fei et al., 2024). Specifically, researchers adopt two main strategies:

- **Position-Aware Data Design.** This strategy involves structuring data to force uniform attention across the entire sequence. For example, PAM-QA (He et al., 2024) decomposes multi-document data to direct the model's attention. Similarly, FILM-7B (An et al., 2024c) generates question-answer pairs with randomly positioned key evidence. Extending this to multimodal domains, embedding explicit page indices into training significantly enhances retrieval performance for long visual documents (Veselka, 2026).
- **Explicit Structured Data Design.** This approach trains models on explicitly structured data. Methods include extracting knowledge graph triples (Baker et al., 2024) or creating synthetic key-value datasets (Xiong et al., 2025). Training on these explicit structures improves retrieval accuracy and reduces hallucinations.

Retrieval evaluation datasets typically follow two main paradigms:

- **Synthetic Needle-Insertion Datasets.** A primary example is NIAH, which tests single-key retrieval from irrelevant background text (gkamradt, 2023). Building on this, NeedleBench (Li et al., 2024b) constructs multi-needle tasks, RULER (Hsieh et al., 2024) explores performance under diverse noise conditions, and Counting-Stars (Song et al., 2025b) measures how models handle overlapping dependencies. However, early synthetic needle tests often suffer from an over-reliance on exact keyword matching (Du et al., 2023; Modarressi et al., 2025b).
- **Retrieval within Complex Tasks.** Because information localization frequently overlaps with other capabilities, retrieval is often evaluated as a core component of broader tasks rather than an isolated skill. Datasets for complex tasks, such as FanoutQA (Zhu et al., 2024) and Loong (Wang et al., 2024b), naturally test retrieval by requiring models to locate key evidence within long contexts before reasoning (Li et al., 2025b; Ni et al., 2024; Laban et al., 2024). Additionally, multi-task benchmarks like Infty-Bench (Zhang et al., 2024b) and HELMET (Yen et al., 2025) design specific sub-tasks to measure retrieval accuracy within broader application scenarios.

Summary.

- **Training.** Recent works indicates that position-aware and explicitly structured data can effectively mitigate attention biases like the "lost in the middle".
- **Evaluation.** Benchmarks are shifting from simple matching in synthetic contexts toward complex semantic retrieval to prevent shortcut learning.

4.3 Reasoning Capability

The reasoning capability of LCMs involves logical deduction by synthesizing distant contextual information with internal knowledge (Wan et al., 2025a; Yang et al., 2025d). This enables them to understand complex logical structures like causal and conditional relationships, and thereby address tasks requiring deep analysis (Ding et al., 2025).

Models often exhibit weaker reasoning over long contexts compared to short ones (GLM et al., 2024; Liu et al., 2024e). To bridge this gap, recent research prioritizes constructing synthetic data that explicitly models the reasoning process (Zhang et al., 2024c; Tang et al., 2025). These synthesis strategies generally fall into three paradigms:

- **Explicit Reasoning Paths.** Standard QA pairs often obscure reasoning paths; thus, recent works emphasize exposing intermediate steps (Sun et al., 2026b). To achieve this, InftyThink (Yan et al., 2025b) restructures problems iteratively to enforce sequential learning. Concurrently, scalable self-distillation methods, such as Semantic Soft Bootstrapping (Mitra and Ulukus, 2025) and Light-R1 (Wen et al., 2025), allow models to learn from their own reasoning trajectories.
- **Integrating Multi-Hop Evidence.** Reasoning often requires connecting information across sources. Approaches like LongMIT (Chen et al., 2025c) and MDCure (Liu et al., 2025a) guide models to logically combine evidence rather than concatenating it. To keep this process factual, LongFaith (Yang et al., 2025b) applies citation supervision to enforce faithful evidence usage.
- **Constructing Verifiable Reward Signals.** To address sparse supervision in long-context Reinforcement Learning (RL), methods like LongRLVR (Chen et al., 2026a) and LongR (Ping et al., 2026) introduce contextual verification and utility rewards. This provides dense signals that reinforce specific reasoning steps.

Reasoning benchmarks for LCMs have evolved from evaluating basic comprehension to assessing deep logical inference (Amos et al., 2024). While early single-document evaluations often suffered from highly concentrated evidence (Ling et al., 2025; Bai et al., 2025a), recent benchmarks introduce stricter demanding conditions.

- **Complex Evidence Integration.** Evaluating the ability to synthesize disparate information is crucial for long-context reasoning. FanoutQA (Zhu et al., 2024) utilizes Wikipedia's structure for multi-hop questions, and BABILong (Kuratov et al., 2024) incorporates twenty distinct reasoning tasks. Highlighting the necessity of evidence traceability, SIN-Bench (Ren et al., 2026) introduces a "no evidence, no score" metric for long scientific literature, revealing that models tend to guess answers based on memory.
- **Noise Robustness and Memory Management.** To assess stability, Michelangelo (Vodrahalli et al., 2024) challenges models to maintain reasoning chains in noisy contexts. For reward models, MemoryRewardBench (Tang et al., 2026) evaluates long-term memory and coherent generation across contexts up to 128K tokens.
- **Domain-Specific Reasoning.** To assess practical applicability, recent benchmarks target specialized domains such as programming (LongCodeBench (Rando et al., 2025)), finance (DocFinQA (Reddy et al., 2024), FinTextQA (Chen et al., 2024a)), and clinical healthcare (LongHealth (Adams et al., 2025)).

Summary.

- **Training.** Recent works suggests that reasoning significantly benefit from explicit inferential traces, multi-hop evidence integration, and verifiable reward signals.
- **Evaluation.** These trends indicate that reasoning evaluation demands traceable evidence synthesis, noise robustness, and domain-specific application.

4.4 Aggregation Capability

The aggregation capability of LCMs involves capturing dependencies across extensive contexts to synthesize and summarize distributed information (Li et al., 2025c; Yang et al., 2026). While reasoning focuses on sequential logical deduction, aggregation emphasizes information compression and global pattern recognition (Li et al., 2025b).

To train this capability, data strategies must emphasize global dependencies and noise filtering. Recent approaches broadly fall into three categories:

- **Modeling Noise and Distractors.** Since effective information compression requires discarding irrelevant noise (Li et al., 2024a), NExtLong (Gao et al., 2025c) augments training data with explicit distractors, forcing models to distinguish relevant contexts. Other methods control the spatial distribution of key information to train global attention (Tang et al., 2025).
- **Structuring Global Dependencies.** To make long-range dependencies more learnable, LongABC (Wu et al., 2025a) reformats training objectives to help self-attention quantify distant relationships. SkipAlign (Wu et al., 2024) leverages explicit positional indexing to map semantic structures. Furthermore, restructuring standard QA datasets into summarization templates provides stronger supervisory signals (Kryściński et al., 2022; Zhang et al., 2024c).
- **Constructing Preference Data for Faithfulness.** To mitigate hallucinations in long-form summarization, frameworks like SCRPO (Hu et al., 2025a) leverage self-evaluation and refinement to construct preference datasets. Optimizing on these self-generated pairs directly aligns model outputs with the source context.

Evaluation benchmarks for aggregation focus on summarization and cross-document analysis, which frequently overlap with retrieval and reasoning (Koh et al., 2022; Yan et al., 2025a).

- **Progressive and Traceable Synthesis.** Evaluating advanced aggregation requires moving beyond standard summarization to test hierarchical information integration. LCFO (Costa-jussà et al., 2025) specifically benchmarks progressive summarization capabilities. Furthermore, SummHay (Laban et al., 2024) stress-tests under extreme conditions, requiring models to generate coherent summaries from massive, noisy corpora while providing accurate source citations.
- **Statistical and Structural Aggregation.** This involves quantitative and categorical synthesis. Oolong (Bertsch et al., 2025) introduces a specialized benchmark for information analysis and numerical aggregation. HoloBench (Maekawa et al., 2025) evaluates database-style operations over unstructured long contexts.
- **Core Dependency Evaluation.** XL2Bench (Ni et al., 2024) evaluates long-range dependency

aggregation strictly independent of other skills.

Summary.

- **Training.** Developing aggregation capabilities relies on structuring global dependencies, modeling explicit distractors, and applying preference optimization.
- **Evaluation.** Existing evaluations assess aggregation through traceable synthesis, complex statistical operations, and isolated dependency testing in noisy.

4.5 Long-Form Generation Capability

The long-form generation capability refers to the generation of coherent, fluent, and logically structured extensive content, such as articles and reports (Li et al., 2023; Wan et al., 2025b; Ye et al., 2025b). While previously discussed capabilities primarily focus on how LCMs comprehend extensive input contexts, long-form generation emphasizes the capacity to produce extended outputs.

Standard models struggle with long-form generation because next token prediction lacks long-range planning (Sun et al., 2026a). To overcome this, data-centric approaches focus on teaching models structural organization. Recent training strategies generally follow three directions:

- **Creating Planning Data.** This direction embeds explicit planning into training data. For example, DeFine (Wang et al., 2025c) and DOME (Wang et al., 2025d) augment data with fine-grained plans and narrative structures, while DeepWriting (Wang et al., 2025a) creates comprehensive planning data by reverse-engineering reasoning trajectories.
- **Imitating Dynamic Generation Processes.** To simulate human writing workflows, recent methods train models on multi-step pipelines. LongWriter (Bai et al., 2025b) utilizes a sequential plan-then-generate pipeline, whereas SuperWriter (Wu et al., 2025b) introduces an iterative plan-write-refine loop to enable self-correction.
- **Constructing Preference and Reward Data.** Aligning models with global structural quality requires specialized feedback data. To generate reliable reward signals for extensive outputs, frameworks like OpenReward (Hu et al., 2025b), ProxyReward (Guo et al., 2025), and ACE-RL (Chen et al., 2025a) synthesize automated signals via tools, structured prompts, or fine-grained constraints. Furthermore, LongDPO (Ping et al., 2025) and Hierarchical DPO (Wang et al., 2025d) construct step-by-step preference data to provide targeted feedback on planning stages.

Since traditional n-gram metrics like ROUGE fail to capture structural integrity and logical flow, recent benchmarks increasingly use LLM-as-a-judge and specialized frameworks to evaluate long-form generation across three critical dimensions:

- **Verifiability and Factual Accuracy.** To measure semantic utility and factual correctness, ProxyQA (Tan et al., 2024) evaluates information coverage through question-answering formats, while LongFact (Wei et al., 2024) utilizes automated search tools for verification. Additionally, LongWeave (Xiao et al., 2025) incorporates constraint verification to balance real-world complexity with objective scoring.
- **Structural Coherence and Process Consistency.** To evaluate whether models maintain logical soundness during generation (Hsu et al., 2025). Benchmarks such as LongLaMP (Kumar et al., 2024), RAPID (Gu et al., 2025), and LongGenBench (Liu et al., 2024c) test whether the output strictly adheres to its initial writing plan and preserves structural consistency.
- **Practical Application and Stability.** For real-world usage, ExpertLongBench (Ruan et al., 2025) evaluates domain-specific tasks with expert rubrics, while AcademicEval (Zhang et al., 2025a) employs recent academic papers for dynamic evaluation. Furthermore, UNCLE (Yang et al., 2025c) and LongInOutBench (Zhang et al., 2025c) quantify uncertainty expression and context preservation stability during writing tasks.

Summary.

- **Training.** To enhance long-form generation, current strategies focus on constructing hierarchical datasets, simulating human writing workflows, and synthesizing step-by-step preference feedback.
- **Evaluation.** Benchmarks use structured constraints and dynamic rubrics to evaluate factual accuracy, structural coherence, and practical stability.

5 Discussion and Future Directions

In this work, we present a systematic review of data strategies for long-context language models (LCMs), emphasizing the relationship between

data characteristics and model capabilities. Although current works provide useful guidelines for dataset design, several challenges remain. This section highlights these challenges and outlines specific directions for future research.

Data Quality Assessment. A major challenge in training LCMs is the limited availability of high-quality, naturally long documents. While researchers often use LLMs to generate synthetic long texts, ensuring factual consistency and logical coherence across long sequences remains difficult (Deng et al., 2025; Veselka, 2026; Ping et al., 2026). Currently, the field lacks a standard way to define clear criteria for high-quality long-context data. Relying on simple text statistics is usually insufficient for evaluating complex document structures (Sun et al., 2026b). Therefore, future work should develop a unified framework that quantifies data quality across dimensions like information density, structural dependency distance, and cross-document coherence. Establishing such metrics would help automate data selection, significantly reduce the costs of processing large-scale datasets, and improve the reliability of synthetic data.

Data Mixing and Length Distribution. Extending the input length introduces high computational costs. To reduce this overhead, researchers often train models on a mixture of short and long texts. However, the specific effects of these mixing strategies are not fully understood (Zhang et al., 2024a; An et al., 2024b). Training on too many short texts can degrade the model's ability to handle long sequences. To mitigate this, researchers must carefully balance the data distribution, yet finding the optimal mixing ratio currently relies on heavy manual tuning (Salemi et al., 2025; Zhu et al., 2025; Ye et al., 2025a). A valuable future direction is to establish scaling laws specific to context length distributions. For example, researchers should study whether packing multiple short documents together provides the same training benefits as using a single naturally long document. Additionally, exploring curriculum learning methods that gradually increase data length during training could offer a more systematic way to balance training efficiency with long-context capabilities.

Dynamic Evaluation Frameworks. As LCMs improve quickly, they easily achieve high scores on existing evaluation benchmarks. Moreover, current evaluation methods often fail to reflect the complex tasks users actually expect these models to handle (Song et al., 2025a; Wan and Ma, 2025). Since constantly creating new benchmarks by hand is too expensive (Zeng et al., 2026; Chen et al., 2026b), the community needs to shift toward dynamic evaluation frameworks. Future research could explore the automatic generation of test queries where researchers can dynamically adjust the complexity, context length, and required reasoning steps. This flexibility would allow tests to become progressively more challenging as models grow stronger. By designing these adjustable tests around real-world user needs, we can ensure that evaluation always remains challenging and meaningful in revealing the true capabilities of LCMs.

Acknowledgments

We want to thank all the anonymous reviewers for their valuable comments. This work was supported by the National Science Foundation of China (NSFC No. 62576232), the Suzhou Science and Technology Key R&D Project / Suzhou Key Core Technology Project (Project No. SYG2025069), Key Laboratory of Data Intelligence and Advanced Computing, Soochow University.

References

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. 2025. Longhealth: A question answering benchmark with long clinical documents. *Journal of Healthcare Informatics Research*, pages 1–17.

Chandra Vamsi Krishna Alla, Harish Naidu Gaddam, and Manohar Kommi. 2026. Budgetmem: Learning selective memory policies for cost-efficient long-context processing in language models.

Ido Amos, Jonathan Berant, and Ankit Gupta. 2024. Never train from scratch: Fair comparison of long-sequence models requires data-driven priors. In *International Conference on Learning Representations*.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024a. L-eval: Instituting standardized evaluation for long context language models.

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411.

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. 2024b. Why does the effective context length of llms fall short?

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024c. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024a. Longalign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1376–1395.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2024b. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025a. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, Vienna, Austria. Association for Computational Linguistics.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025b. Longwriter: Unleashing 10,000+ word generation from long context llms. In *The Thirteenth International Conference on Learning Representations*.

George Arthur Baker, Ankush Raut, Sagi Shaier, Lawrence E Hunter, and Katharina von der Wense. 2024. Lost in the middle, and in-between: Enhancing language models' ability to reason over long contexts in multi-hop qa.

Amanda Bertsch, Adithya Pratapa, Teruko Mitamura, Graham Neubig, and Matthew R Gormley. 2025. Oolong: Evaluating long context reasoning and aggregation capabilities. *arXiv preprint arXiv:2511.02817*.

Bernd Bohnet, Kevin Swersky, Rosanne Liu, Pranjal Awasthi, Azade Nova, Javier Snaider, Hanie Sedghi, Aaron T Parisi, Michael Collins, Angeliki Lazaridou, et al. 2024. Long-span question-answering: Automatic question generation and qa-system ranking via side-by-side evaluation. *arXiv preprint arXiv:2406.00179*.

Guanzheng Chen, Michael Qizhe Shieh, and Lidong Bing. 2026a. Longrlvr: Long-context reinforcement learning requires verifiable context rewards. *arXiv preprint arXiv:2603.02146*.

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024a. Fintextqa: A dataset for long-form financial question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6025–6047.

Jianghao Chen, Wei Sun, Qixiang Yin, Zhixing Tan, and Jiajun Zhang. 2025a. Ace-rl: Adaptive constraint-enhanced reward for long-form generation reinforcement learning. *arXiv preprint arXiv:2509.04903*.

Jianghao Chen, Junhong Wu, Yangyifan Xu, and Jiajun Zhang. 2025b. LADM: Long-context training data selection with attention-based dependency measurement for LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3076–3090, Vienna, Austria. Association for Computational Linguistics.

Longze Chen, Ziqiang Liu, Wanwei He, Yinhe Zheng, Hao Sun, Yunshui Li, Run Luo, and Min Yang. 2024b. Long context is not long at all: A prospector of long-dependency data for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8222–8234.

Zhi Chen, Qiguang Chen, Libo Qin, Qipeng Guo, Haijun Lv, Yicheng Zou, Hang Yan, Kai Chen, and Dahua Lin. 2025c. What are the essential

factors in crafting effective long context multi-hop instruction datasets? insights and best practices. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27129–27151, Vienna, Austria. Association for Computational Linguistics.

Ziyang Chen, Xing Wu, Junlong Jia, Chaochen Gao, Qi Fu, Debing Zhang, and Songlin Hu. 2026b. Longbench pro: A more realistic and comprehensive bilingual long-context evaluation benchmark. *arXiv preprint arXiv:2601.02872*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Marta R. Costa-jussà, Pierre Andrews, Mariano Coria Meglioli, Joy Chen, Joe Chuang, David Dale, Christophe Ropers, Alexandre Mourachko, Eduardo Sánchez, Holger Schwenk, Tuan A. Tran, Arina Turkatenko, and Carleigh Wood. 2025. LCFO: Long context and long form output dataset and benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10672–10700, Vienna, Austria. Association for Computational Linguistics.

Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Christian Norgaard, Nayantara Mudur, Martyna Beata Plomecka, Paul Raccuglia, et al. 2025. Curie: Evaluating llms on multitask scientific long-context understanding and reasoning. In *The Thirteenth International Conference on Learning Representations*.

Haoran Deng, Yingyu Lin, Zhenghao Lin, Xiao Liu, Yizhou Sun, Yi-An Ma, and Yeyun Gong. 2025. Beyond length: Quantifying long-range information for long-context llm pretraining data. *arXiv preprint arXiv:2510.25804*.

Jiayu Ding, Shuming Ma, Lei Cui, Nanning Zheng, and Furu Wei. 2025. Longreasonarena: A long reasoning benchmark for large language models. *arXiv preprint arXiv:2508.19363*.

Guancheng Du, Yong Hu, Wenqing Wang, Yaming Yang, and Jiaheng Gao. 2025. Sagascale: A realistic, scalable, and high-quality long-context benchmark built from full-length novels. *arXiv preprint arXiv:2601.09723*.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1):110–120.

Weizhi Fei, Xueyan Niu, Guoqing Xie, Yanhua Zhang, Bo Bai, Lei Deng, and Wei Han. 2024. Retrieval meets reasoning: Dynamic in-context editing for long-text understanding. *arXiv preprint arXiv:2406.12331*.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. In *Proceedings of the 41st International Conference on Machine Learning*, pages 14125–14134.

Chaochen Gao, Xing Wu, Zijia Lin, Debing Zhang, and Songlin Hu. 2025a. Longmagpie: A self-synthesis method for generating large-scale long-context instructions. *arXiv preprint arXiv:2505.17134*.

Chaochen Gao, W Xing, Qi Fu, and Songlin Hu. 2025b. Quest: Query-centric data synthesis approach for long-context scaling of large language model. In *The Thirteenth International Conference on Learning Representations*.

Chaochen Gao, W Xing, Zijia Lin, Debing Zhang, and Songlin Hu. 2025c. Nextlong: Toward effective long-context training without long documents. In *Forty-second International Conference on Machine Learning*.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2024. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*.

Yunfan Gao, Yun Xiong, Wenlong Wu, Zijing Huang, Bohan Li, and Haofen Wang. 2025d. Uniah: Unified rag and llm evaluation for long context needle-in-a-haystack. *arXiv preprint arXiv:2503.00353*.

Shawn Gavin, Tuney Zheng, Jiaheng Liu, Quehry Que, Noah Wang, Jian Yang, Chenchen Zhang, Wenhao Huang, Wenhu Chen, and Ge Zhang. 2024. Longins: A challenging long-context instruction-based exam for llms. *arXiv preprint arXiv:2406.17588*.

gkamradt. 2023. Llmtest-needleinahaystack. https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

GLM-5-Team. 2026. Glm-5: from vibe coding to agentic engineering.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. 2024. Is it really long context if all you need is retrieval? towards genuinely difficult long context nlp. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16576–16586.

Hongchao Gu, Dexun Li, Kuicai Dong, Hao Zhang, Hang Lv, Hao Wang, Defu Lian, Yong Liu, and Enhong Chen. 2025. RAPID: Efficient retrieval-augmented long text generation with writing planning and information discovery. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16742–16763, Vienna, Austria. Association for Computational Linguistics.

Zhihan Guo, Jiele Wu, Wenqian Cui, Yifei Zhang, Minda Hu, Yufei Wang, and Irwin King. 2025. From general reward to targeted reward: Improving open-ended long-context generation models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5151–5166.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, LiuYiBo LiuYiBo, Qianguosun Qianguosun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2024. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13628–13642.

Linda He, WANG Jue, Maurice Weber, Shang Zhu, Ben Athiwaratkun, and Ce Zhang. 2025. Scaling instruction-tuned llms to million-token contexts via hierarchical synthetic data generation. In *The Thirteenth International Conference on Learning Representations*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Chi-Yang Hsu, Alexander Braylan, Yiheng Su, Omar Alonso, and Matthew Lease. 2025. Instance-level performance prediction for long-form generation tasks. *arXiv e-prints*, pages arXiv–2509.

Ting-Yao Hu, Hema Swetha Koppula, Hadi Pouransari, Cem Koc, Oncel Tuzel, and Raviteja Vemulapalli. 2025a. Learning from self critique and refinement for faithful llm summarization. *arXiv preprint arXiv:2512.05387*.

Ziyou Hu, Zhengliang Shi, Minghang Zhu, Haitao Li, Teng Sun, Pengjie Ren, Suzan Verberne, and Zhaochun Ren. 2025b. Openreward: Learning to reward long-form agentic tasks via reinforcement learning. *arXiv preprint arXiv:2510.24636*.

Junlong Jia, Xing Wu, Chaochen Gao, Ziyang Chen, Zijia Lin, Zhongzhi Li, Weinong Wang, Haotian Xu, Donghui Jin, Debing Zhang, et al. 2025. Litelong: Resource-efficient long-context data synthesis for llms. *arXiv preprint arXiv:2509.15568*.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Fables: Evaluating faithfulness and content selection in book-length summarization. In *First Conference on Language Modeling*.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM computing surveys*, 55(8):1–35.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. Booksum: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: testing the limits of llms with long context reasoning-in-a-haystack. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 106519–106554.

Philippe Laban, Alexander Richard Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and rag systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9885–9903.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Xuanyu Lei, Chenliang Li, Yuning Wu, Kaiming Liu, Weizhou Shen, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. 2025. Writing-rl: Advancing long-form writing via adaptive curriculum reinforcement learning. *arXiv preprint arXiv:2506.05760*.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How long can context length of open-source llms truly promise? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024a. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333.

Jiaxi Li, Xingxing Zhang, Xun Wang, Xiaolong Huang, Li Dong, Liang Wang, Si-Qing Chen, Wei Lu, and Furu Wei. 2025a. Wildlong: Synthesizing realistic long-context instruction data at scale. *arXiv preprint arXiv:2502.16684*.

Kuan Li, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang, and Minhao Cheng. 2025b. Lara: Benchmarking retrieval-augmented generation and long-context llms–no silver bullet for lc or rag routing. In *Forty-second International Conference on Machine Learning*.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. 2024b. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *arXiv preprint arXiv:2407.11963*.

Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. 2024c. Large language models can self-improve in long-context reasoning. *arXiv preprint arXiv:2411.08147*.

Taiji Li, Hao Chen, Fei Yu, and Yin Zhang. 2025c. Hera: Improving long document summarization using large language models with context packaging and reordering.

Wenhao Li, Bangcheng Sun, Weihao Ye, Tianyi Zhang, Daohai Yu, Fei Chao, and Rongrong Ji. 2026. Ccf: A context compression framework for efficient long-sequence language modeling.

Zongxia Li, Yapei Chang, Yuhang Zhou, Xiyang Wu, Zichao Liang, Yoo Yeon Sung, and Jordan Lee Boyd-Graber. 2025d. Semantically-aware rewards for open-ended r1 training in free-form generation. *arXiv preprint arXiv:2506.15068*.

Zhan Ling, Kang Liu, Kai Yan, Yifan Yang, Weijian Lin, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. 2025. Longreason: A synthetic long-context reasoning benchmark via context expansion. *arXiv preprint arXiv:2501.15089*.

Gabrielle Kaili-May Liu, Bowen Shi, Avi Caciularu, Idan Szpektor, and Arman Cohan. 2025a. MDCure: A scalable pipeline for multi-document instruction-following. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29258–29296, Vienna, Austria. Association for Computational Linguistics.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, Yuanxing Zhang, Zhuo Chen, Hangyu Guo, Shilong Li, Ziqiang Liu, Yong Shan, Yifan Song, Jiayi Tian, Wenhao Wu, Zhejian Zhou, Ruijie Zhu, Junlan Feng, Yang Gao, Shizhu He, Zhoujun Li, Tianyu Liu, Fanyu Meng, Wenbo Su, Yingshui Tan, Zili Wang, Jian Yang, Wei Ye, Bo Zheng, Wangchunshu Zhou, Wenhao Huang, Sujian Li, and Zhaoxiang Zhang. 2025b. A comprehensive survey on long context language modeling.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. 2024b. Regmix: Data mixture as regression for language model pre-training. In *The Thirteenth International Conference on Learning Representations*.

Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. 2024c. Longgenbench: Long-context generation benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 865–883.

Xiaoran Liu, Kai Lv, Qipeng Guo, Hang Yan, Conghui He, Xipeng Qiu, and Dahua Lin. 2024d. LongWanjuan: Towards systematic measurement for long text quality. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5709–5725, Miami, Florida, USA. Association for Computational Linguistics.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024e. Chatqa: Building gpt-4 level conversational qa models. *CoRR*.

Yida Lu, Jiale Cheng, Zhexin Zhang, Shiyao Cui, Cunxiang Wang, Xiaotao Gu, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. 2025. Longsafety: Evaluating long-context safety of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31705–31725.

Seiji Maekawa, Hayate Iso, and Nikita Bhutani. 2025. Holistic reasoning with long-context lms: A benchmark for database operations on massive textual data. In *The Thirteenth International Conference on Learning Representations*.

Mounika Marreddy, Subba Reddy Oota, Venkata Charan Chinni, Manish Gupta, and Lucie Flek. 2025. USDC: A dataset of *U*ser *S*tance and *D*ogmatism in long *C*onversations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23715–23759, Vienna, Austria. Association for Computational Linguistics.

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, Enwei Jiao, Gengxin Li, Guojun Zhang, Haohai Sun, Houze Dong, Jiadai Zhu, Jiaqi Zhuang, Jiayuan Song, Jin Zhu, Jingtao Han, Jingyang Li, Junbin Xie, Junhao Xu, Junjie Yan, Kaishun Zhang, Kecheng Xiao, Kexi Kang, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Zheng, Linbo Chai, Long Xing, Meizhi Ju, Mingyuan Chi, Mozhi Zhang, Peikai Huang, Pengcheng Niu, Pengfei Li, Pengyu Zhao, Qi Yang, Qidi Xu, Qiexiang Wang, Qin Wang, Qiuhui Li, Ruitao Leng, Shengmin Shi, Shuqi Yu, Sichen Li, Songquan Zhu, Tao Huang, Tianrun Liang, Weigao Sun, Weixuan Sun, Weiyu Cheng, Wenkai Li, Xiangjun Song, Xiao Su, Xiaodong Han, Xinjie Zhang, Xinzhu Hou, Xu Min, Xun Zou, Xuyang Shen, Yan Gong, Yingjie Zhu, Yipeng Zhou, Yiran Zhong, Yongyi Hu, Yuanxiang Fan, Yue Yu, Yufeng Yang, Yuhao Li, Yunan Huang, Yunji Li, Yunpeng Huang, Yunzhi Xu, Yuxin Mao, Zehan Li, Zekang Li, Zewei Tao, Zewen Ying, Zhaoyang Cong, Zhen Qin, Zhenhua Fan, Zhihang Yu, Zhuo Jiang, and Zijia Wu. 2025. Minimax-01:

Scaling foundation models with lightning attention.

Purbesh Mitra and Sennur Ulukus. 2025. Semantic soft bootstrapping: Long context reasoning in llms without reinforcement learning. *arXiv preprint arXiv:2512.05105*.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A Rossi, Seunghyun Yoon, and Hinrich Schuetze. 2025a. Nolima: Long-context evaluation beyond literal matching. In *Forty-second International Conference on Machine Learning*.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A Rossi, Seunghyun Yoon, and Hinrich Schuetze. 2025b. Nolima: Long-context evaluation beyond literal matching. In *International Conference on Machine Learning*, pages 44554–44570. PMLR.

Xuanfan Ni, Hengyi Cai, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, and Piji Li. 2024. Xl^{2}bench: A benchmark for extremely long context understanding with long-range dependencies.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Miao Peng, Weizhou Shen, Nuo Chen, Chenliang Li, Ming Yan, and Jia Li. 2026. Incentivizing in-depth reasoning over long contexts with process advantage shaping. *arXiv preprint arXiv:2601.12465*.

Bowen Ping, Zijun Chen, Yiyao Yu, Tingfeng Hui, Junchi Yan, and Baobao Chang. 2026. Longr: Unleashing long-context reasoning via reinforcement learning with dense utility rewards. *arXiv preprint arXiv:2602.05758*.

Bowen Ping, Jiali Zeng, Fandong Meng, Shuo Wang, Jie Zhou, and Shanghang Zhang. 2025. LongDPO: Unlock better long-form generation abilities for LLMs via critique-augmented stepwise information. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7613–7632, Vienna, Austria. Association for Computational Linguistics.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, et al. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *arXiv preprint arXiv:2409.16191*.

Stefano Rando, Luca Romani, Alessio Sampieri, Yuta Kyuragi, Luca Franco, Fabio Galasso, Tatsunori Hashimoto, and John Yang. 2025. Longcodebench: Evaluating coding llms at 1m context windows. *arXiv preprint arXiv:2505.07897*.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458.

Yiming Ren, Junjie Wang, Yuxin Meng, Yihang Shi, Zhiqiang Lin, Ruihang Chu, Yiran Xu, Ziming Li, Yunfei Zhao, Zihan Wang, et al. 2026. Sin-bench: Tracing native evidence chains in long-context multimodal scientific interleaved literature. *arXiv preprint arXiv:2601.10108*.

Jonathan Roberts, Kai Han, and Samuel Albanie. 2025. Needle threading: Can llms follow threads through near-million-scale haystacks? In *The Thirteenth International Conference on Learning Representations*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Jie Ruan, Inderjeet Nair, Shuyang Cao, Amy Liu, Sheza Munir, Micah Pollens-Dempsey, Tiffany Chiang, Lucy Kates, Nicholas David, Sihan Chen, et al. 2025. Expertlongbench: Benchmarking language models on expert-level long-form generation tasks with structured checklists. *arXiv preprint arXiv:2506.01241*.

Alireza Salemi, Julian Killingback, and Hamed Zamani. 2025. Expert: Effective and explainable evaluation of personalized long-form text generation.

Alfred Shen and Aaron Shen. 2026. Gated sparse attention: Combining computational efficiency

with training stability for long-context language models.

Weizhou Shen, Ziyi Yang, Chenliang Li, Zhiyuan Lu, Miao Peng, Huashan Sun, Yingcheng Shi, Shengyi Liao, Shaopeng Lai, Bo Zhang, et al. 2025. Qwenlong-l1. 5: Post-training recipe for long-context reasoning and memory management. *arXiv preprint arXiv:2512.12967*.

Dachuan Shi, Yonggan Fu, Xiangchi Yuan, Zhongzhi Yu, Haoran You, Sixu Li, Xin Dong, Jan Kautz, Pavlo Molchanov, Yingyan, and Lin. 2025. Lacache: Ladder-shaped kv caching for efficient long-context modeling of large language models.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Xi Victoria Lin, Noah A Smith, Luke Zettlemoyer, Wen-tau Yih, and Mike Lewis. 2024. In-context pretraining: Language modeling beyond document boundaries. In *The Twelfth International Conference on Learning Representations*.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025a. PRMBench: A fine-grained and challenging benchmark for process-level reward models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25299–25346, Vienna, Austria. Association for Computational Linguistics.

Mingyang Song, Mao Zheng, and Xuan Luo. 2025b. Counting-stars: A multi-evidence, position-aware, and scalable benchmark for evaluating long-context large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3753–3763, Abu Dhabi, UAE. Association for Computational Linguistics.

Mingyang Song, Mao Zheng, and Xuan Luo. 2025c. Counting-stars: A multi-evidence, position-aware, and scalable benchmark for evaluating long-context large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3753–3763.

Konrad Staniszewski, Szymon Tworkowski, Sebastian Jaszczur, Yu Zhao, Henryk Michalewski, Łukasz Kuciński, and Piotr Miłoś. 2025. Structured packing in llm training improves long context utilization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25201–25209.

Zechen Sun, Yuyang Sun, Zecheng Tang, Juntao Li, Wenpeng Hu, Wenliang Chen, Zhunchen Luo, Guotong Geng, and Min Zhang. 2026a. Is-cot: Breaking the long-form generation collapse via interleaved structural thinking.

Zechen Sun, Zecheng Tang, Juntao Li, Wenpeng Hu, Wenliang Chen, Zhunchen Luo, and Qiaoming Zhu. 2026b. Cgmis: Concept-graph based multi-hop instructions synthesis for enhancing long-context reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33153–33161.

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, et al. 2024. Proxyqa: An alternative framework for evaluating long-form text generation with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6806–6827.

Zecheng Tang, Baibei Ji, Ruoxi Sun, Haitian Wang, WangJie You, Zhang Yijun, Wenpeng Zhu, Ji Qi, Juntao Li, and Min Zhang. 2026. Memoryrewardbench: Benchmarking reward models for long-term memory management in large language models. *arXiv preprint arXiv:2601.11969*.

Zecheng Tang, Zechen Sun, Juntao Li, Qiaoming Zhu, and Min Zhang. 2025. Logo—long context alignment via efficient preference optimization. In *Forty-second International Conference on Machine Learning*.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei

Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. 2026. Kimi k2.5: Visual agentic intelligence.

Junfeng Tian, Da Zheng, Yang Chen, Rui Wang, Colin Zhang, and Debing Zhang. 2025. Untie the knots: An efficient data augmentation strategy for long-context pre-training in language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1223–1242, Vienna, Austria. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Austin Veselka. 2026. How to train your long-context visual document model. *arXiv preprint arXiv:2602.15257*.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, et al. 2024. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *arXiv preprint arXiv:2409.12640*.

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. 2025a. Qwenlong-l1: Towards long-context large reasoning models with reinforcement learning.

Kaiyang Wan, Honglin Mu, Rui Hao, Haoran Luo, Tianle Gu, and Xiuying Chen. 2025b. A cognitive writing perspective for constrained long-form text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9832–9844, Vienna, Austria. Association for Computational Linguistics.

Luanbo Wan and Weizhi Ma. 2025. Storybench: A dynamic benchmark for evaluating long-term memory with multi turns. *arXiv preprint arXiv:2506.13356*.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Qian Wang, and Yue Zhang. 2024a. Novelqa: A benchmark for long-range novel question answering. *arXiv e-prints*, pages arXiv–2403.

Haozhe Wang, Haoran Que, Qixin Xu, Minghao Liu, Wangchunshu Zhou, Jiazhan Feng, Wanjun Zhong, Wei Ye, Tong Yang, Wenhao Huang, et al. 2025a. Reverse-engineered reasoning for open-ended generation. *arXiv preprint arXiv:2509.06160*.

Lanrui Wang, Mingyu Zheng, Hongyin Tang, Zheng Lin, Yanan Cao, Jingang Wang, Xunliang Cai, and Weiping Wang. 2025b. Needleinatable: Exploring long-context capability of large language models towards long-structured tables.

Ming Wang, Fang Wang, Minghao Hu, Li He, Haiyang Wang, Jun Zhang, Tianwei Yan, Li Li, Zhunchen Luo, Wei Luo, Xiaoying Bai, and Guotong Geng. 2025c. Define: A decomposed and fine-grained annotated dataset for long-form article generation.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. 2024b. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *EMNLP*.

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, Daiyuan Li, Yu Hu, and Mingkui Tan. 2025d. Generating long-form story using dynamic hierarchical outlining with memory-enhancement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1352–1391.

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. 2024c. Beyond the limits: A survey of techniques to extend the context length in large language models. In *IJCAI*.

Zhchao Wang, Bin Bi, Yanqi Luo, Sitaram Asur, and Claire Na Cheng. 2025e. Diversity enhances an llm's performance in rag and long-context task. *arXiv preprint arXiv:2502.09017*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. 2024. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Tanglifu Tanglifu, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. 2025. Light-r1: Curriculum SFT, DPO and RL for long COT from scratch and beyond. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 318–327, Vienna, Austria. Association for Computational Linguistics.

Chen Wu and Yin Song. 2025. Scaling context, not parameters: Training a compact 7B language

model for efficient long-context processing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 61–68, Vienna, Austria. Association for Computational Linguistics.

Longyun Wu, Dawei Zhu, Guangxiang Zhao, Zhuocheng Yu, Junfeng Ran, Xiangyu Wong, Lin Sun, and Sujian Li. 2025a. LongAttn: Selecting long-context training data via token-level attention. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19367–19380, Vienna, Austria. Association for Computational Linguistics.

Wenhao Wu, Yizhong Wang, Yao Fu, Xiang Yue, Dawei Zhu, and Sujian Li. 2024. Long context alignment with short instructions and synthesized positions. *arXiv preprint arXiv:2405.03939*.

Yuhao Wu, Yushi Bai, Zhiqiang Hu, Juanzi Li, and Roy Ka-Wei Lee. 2025b. Superwriter: Reflection-driven long-form generation with large language models. *arXiv preprint arXiv:2506.04180*.

Yao Xiao, Lei Wang, Yue Deng, Guanzheng Chen, Ziqi Jin, Jung-jae Kim, Xiaoli Li, Roy Ka-wei Lee, and Lidong Bing. 2026. Document reconstruction unlocks scalable long-context rlvr. *arXiv preprint arXiv:2602.08237*.

Zikai Xiao, Fei Huang, Jianhong Tu, Jianhui Wei, Wen Ma, Yuxuan Zhou, Jian Wu, Bowen Yu, Zuozhu Liu, and Junyang Lin. 2025. LongWeave: A long-form generation benchmark bridging real-world relevance and verifiability. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10386–10417, Suzhou, China. Association for Computational Linguistics.

Zheyang Xiong, Vasilis Papageorgiou, Kangwook Lee, and Dimitris Papailiopoulos. 2025. From artificial needles to real haystacks: Improving retrieval capabilities in llms by finetuning on synthetic data. In *The Thirteenth International Conference on Learning Representations*.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities. In *The Thirteenth International Conference on Learning Representations*.

Xiaoyue Xu, Qinyuan Ye, and Xiang Ren. 2024. Stress-testing long-context language models with lifelong icl and task haystack. *Advances in Neural Information Processing Systems*, 37:15801–15840.

Kai Yan, Zhan Ling, Kang Liu, Yifan Yang, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. 2025a. Mir-bench: Benchmarking llm's long-context intelligence via many-shot in-context inductive reasoning. *arXiv preprint arXiv:2502.09933*.

Yuchen Yan, Yongliang Shen, Yang Liu, Jin Jiang, Mengdi Zhang, Jian Shao, and Yueting Zhuang. 2025b. Inftythink: Breaking the length limits of long-context reasoning in large language models. *arXiv preprint arXiv:2503.06692*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Shengjie Ma, Aofan Liu, Hui Xiong, and Jian Guo. 2025b. LongFaith: Enhancing long-context reasoning in LLMs with faithful synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3236–3256, Vienna, Austria. Association for Computational Linguistics.

Joonho Yang, Seunghyun Yoon, Hwan Chang, Byeongjeong Kim, and Hwanhee Lee. 2026. Hallucinate at the last in long response generation: A case study on long document summarization.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Sen Yang, Nigel Collier, Dong Yu, and Deqing Yang. 2024b. Logu: Long-form generation with uncertainty expressions.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Dong Yu, Nigel Collier, and Deqing Yang. 2025c. Uncle: Uncertainty expressions in long-form generation. *arXiv preprint arXiv:2505.16922*.

Wang Yang, Zirui Liu, Hongye Jin, Qingyu Yin, Vipin Chaudhary, and Xiaotian Han. 2025d. Longer context, deeper thinking: Uncovering the role of long-context ability in reasoning.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. 2025a. Data mixing laws: Optimizing data mixtures by predicting language modeling performance.

Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. 2025b. Longproc: Benchmarking long-context language models on long procedural generation.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2025. Helmet: How to evaluate long-context language models effectively and thoroughly. International Conference on Learning Representations (ICLR).

Tao Yuan, Xuefei Ning, Dong Zhou, Zhijie Yang, Shiyao Li, Minghui Zhuang, Zheyue Tan, Zhuyu Yao, Dahua Lin, Boxun Li, Guohao Dai, Shengen Yan, and Yu Wang. 2024. Lv-eval: A balanced long-context benchmark with 5 length levels up to 256k.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Weihao Zeng, Yuzhen Huang, and Junxian He. 2026. Loca-bench: Benchmarking language agents under controllable and extreme context growth. *arXiv preprint arXiv:2602.07962*.

Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, Raven Yuan, Tuney Zheng, Wei Pang, Xinrun Du, Yiming Liang, Yinghao Ma, Yizhi Li, Ziyang Ma, Bill Lin, Emmanouil Benetos, Huan Yang, Junting Zhou, Kaijing Ma, Minghao Liu, Morry Niu, Noah Wang, Quehry Que, Ruibo Liu, Sine Liu, Shawn Guo, Soren Gao, Wangchunshu Zhou, Xinyue Zhang, Yizhi Zhou, Yubo Wang, Yuelin Bai, Yuhan Zhang, Yuxiang Zhang, Zenith Wang, Zhenzhu Yang, Zijian Zhao, Jiajun Zhang, Wanli Ouyang, Wenhao Huang, and Wenhu Chen. 2024a. Map-neo: Highly capable and transparent bilingual large language model series.

Haozhen Zhang, Tao Feng, Pengrui Han, and Jiaxuan You. 2025a. Academiceval: Live long-context llm benchmark. *Transactions on Machine Learning Research*, 2025:1–32.

Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2025b. LongReward: Improving long-context large language models with AI feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3718–3739, Vienna, Austria. Association for Computational Linguistics.

Junhao Zhang, Richong Zhang, Fanshuang Kong, Ziyang Miao, Yanhan Ye, and Yaowei Zheng. 2025c. Lost-in-the-middle in long-text generation: Synthetic dataset, evaluation framework, and mitigation. *arXiv preprint arXiv:2503.06868*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024b. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand. Association for Computational Linguistics.

Yikai Zhang, Junlong Li, and Pengfei Liu. 2024c. Extending llms' context window with 100 samples. *arXiv preprint arXiv:2401.07004*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-

hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37.

Wenhao Zhu, Pinzhen Chen, Hanxu Hu, Shujian Huang, Fei Yuan, Jiajun Chen, and Alexandra Birch. 2025. Generalizing from short to long: Effective data synthesis for long-context instruction tuning. *arXiv preprint arXiv:2502.15592*.

Yonghao Zhuang, Lanxiang Hu, Longfei Yun, Souvik Kundu, Zhengzhong Liu, Eric P Xing, and Hao Zhang. 2025. Scaling long context training data by long-distance referrals. In *13th International Conference on Learning Representations, ICLR 2025*, pages 81355–81372. International Conference on Learning Representations, ICLR.