Data Foundations of Long-Context Language Models: A Survey

Zechen Sun[♡], **Yuyang Sun**[♡], **Zhaochen Su**[♡], **Zecheng Tang**[♡], **Juntao Li**[♡][*], **Wenliang Chen**[♡]
[♡]Soochow University, SuZhou, China

zcsuns@stu.suda.edu.cn; yysun0799@163.com
suzhaochen0110@gmail.com; zctang@stu.suda.edu.cn
{ljt,wlchen}@suda.edu.cn

Abstract

Long-context large language models (LCMs) have emerged as a focal point of research in natural language processing due to their ability to process ultra-long text sequences. Among the key factors influencing their performance, long-context data serve as a core foundation for LCMs' modeling and differ significantly from traditional short-context data in terms of length and structural complexity. Despite extensive efforts devoted to optimizing or evaluating LCMs through data-centric approaches, there remains a lack of unified comparisons across existing datasets and an insufficient discussion of the essential data characteristics required by LCMs. To address this gap, this survey presents the first systematic review of LCMs from a data-centric perspective. Specifically: (1) We identify the critical characteristics that long-context data should possess; (2) We comprehensively summarize the sources and construction methods of long-context data; (3) We analyze the relationship between long-context data and LCMs' capabilities; and (4) We outline the main challenges faced by LCMs at the data level and provide insights into future research directions. By offering a systematic review, this work aims to provide clear guidelines for data selection and design in LCMs, thereby promoting continued advancements in the field.

1 Introduction

Long-Context Language Models (LCMs) have emerged as a focal point of research in natural language processing, driven by their ability to process ultra-long text sequences (Anthropic; Achiam et al., 2023). These models excel in tasks requiring deep contextual understanding and sustained coherence, such as long-document summarization, code generation, and multi-hop reasoning (Bai et al., 2023; Bolotova-Baranova et al., 2023; Wang et al., 2024a). However, achieving high performance in these tasks depends not only on advancements in model architecture but also critically on the availability and quality of long-context data (Gao et al., 2024b; Chen et al., 2024c). Unlike traditional short-context data, long-context data exhibits unique characteristics—such as extended dependencies, structural complexity, and non-uniform information density—that pose significant challenges for both data collection and model training (Fu et al., 2024; Chen et al., 2024b).

Despite extensive efforts to evaluate and optimize LCMs, there remains a notable gap in the systematic understanding of long-context data (Hsieh et al., 2024; Kuratov et al., 2024). Existing datasets vary widely in terms of quality, diversity, and applicability, with many failing to meet the demands of training robust LCMs (Liu et al., 2024c; Wen et al., 2025). High-quality, naturally occurring long-context corpora remain scarce, while synthetic datasets, though promising, often struggle to achieve an ideal balance between realism and scalability (Liu et al., 2025a). Similarly, current benchmarks for evaluating LCMs are numerous but limited in scenario coverage, making it difficult to comprehensively assess model capabilities across diverse long-context tasks (Costa-jussà et al., 2024; Yan et al., 2025a). As a result, the field faces a pressing need for unified comparisons of existing datasets and a deeper exploration of the data characteristics essential for advancing LCMs.

1.1 Structure of the Survey

In this survey, we provides the systematically review of LCMs from a data-centric perspectiv. The taxonomy of long-context data is illustrated in Figure 1. And Table 1 summarizes and compares long-context datasets from the perspectives of usage stage, source, capability, and construction method. The survey isorganized as follows:

[*] Juntao Li is the Corresponding Author.

1. Section 2 analyzes the critical characteristics that long-context data should possess to effectively support LCMs training and evaluation.
2. Section 3 summarizes the sources and construction methods of long-context data.
3. Section 4 investigates the relationship between long-context data and LCMs capabilities, highlighting how data properties influence model performance.
4. Finally, we identify key challenges at the data level and propose future research directions to address these issues in Section 5.

2 Requirements for Long-Context Data

Simply requiring long-context data to meet a certain length threshold is far from sufficient for effectively developing the capabilities of LCMs (Hu et al., 2024). It is crucial to explore data characteristics that are more conducive to LCMs. Below, we introduce this exploration from two perspectives: training (§2.1) and evaluation (§2.2).

2.1 Training

Contextual Coherence One of the core characteristics of long-context is their essential contextual coherence (Liu et al., 2024c). Training data must absolutely maintain complete semantic coherence to avoid contextual fragmentation caused by data segmentation or concatenation. This helps LLMs better understand complex semantic relationships within long-context and enhances its generation and reasoning capabilities.

Long-Range Dependencies Long-context often contain long-range dependencies that span multiple sentences or even paragraphs (Chen et al., 2024b). The training data should adequately reflect these long-range dependencies, enabling LLMs to learn how to capture and utilize such relationships within long-context. This improves the model's ability to comprehend complex structures.

Cross-Domain Diversity To ensure the model possesses strong generalization capabilities, long-context training data should cover a wide range of domains (e.g., science, literature, news) and genres (e.g., academic papers, novels, reports). Diverse data sources ensure that the model can adapt to various types of long-context tasks, and the mixture proportions of data from each field crucially impact the competence of outcome models (Fu et al., 2024; Liu et al., 2025b; Ye et al., 2025).

Utilization of Structured Data Some long-context data exhibits structured features, such as code, tables, or lists. Proper utilization of this structured information can enhance the model's understanding of long-context, enabling it to process complex document structures more effectively (Pham et al., 2024; Staniszewski et al., 2025).

2.2 Evaluation

The core idea behind the design of evaluation data for LCMs is that "models with longer input contexts should be capable of completing tasks that were previously difficult or impossible to achieve." Effective evaluation data not only ensures the reliability and consistency of models in real-world applications but also provides guidance for model optimization and selection. Therefore, compared to long-context training data, the construction of evaluation data imposes higher requirements (Yan et al., 2025a; Que et al., 2024).

Length Coverage Long-context test data should encompass contexts of varying lengths, with the quantity and quality of evaluation data across different length intervals being as balanced as possible (Yuan et al., 2024). This ensures a flexible assessment of LCMs capabilities in handling long-contexts of varying scales.

Data Authenticity These data should closely resemble real-world scenarios and include various phenomena found in natural language usage, such as colloquial expressions and slang, to ensure that the test reflects the LCMs performance in practical applications (Reddy et al., 2024; Bai et al., 2025).

Uniform Answer Distribution Long-contexts contain vast amounts of information, imposing higher requirements on the annotation and construction of test data (Zhu et al., 2024). The ground truth should not be concentrated in a fixed position within the data, and to avoid LCMs exploiting shortcut learning for predictions (Zhang et al., 2025).

Controlled Difficulty Test data should have moderate difficulty, aligning with the development of LCMs capabilities (Kuratov et al., 2024; Xu et al., 2024b). Avoid excessively difficult or overly simplistic evaluation data to effectively assess the

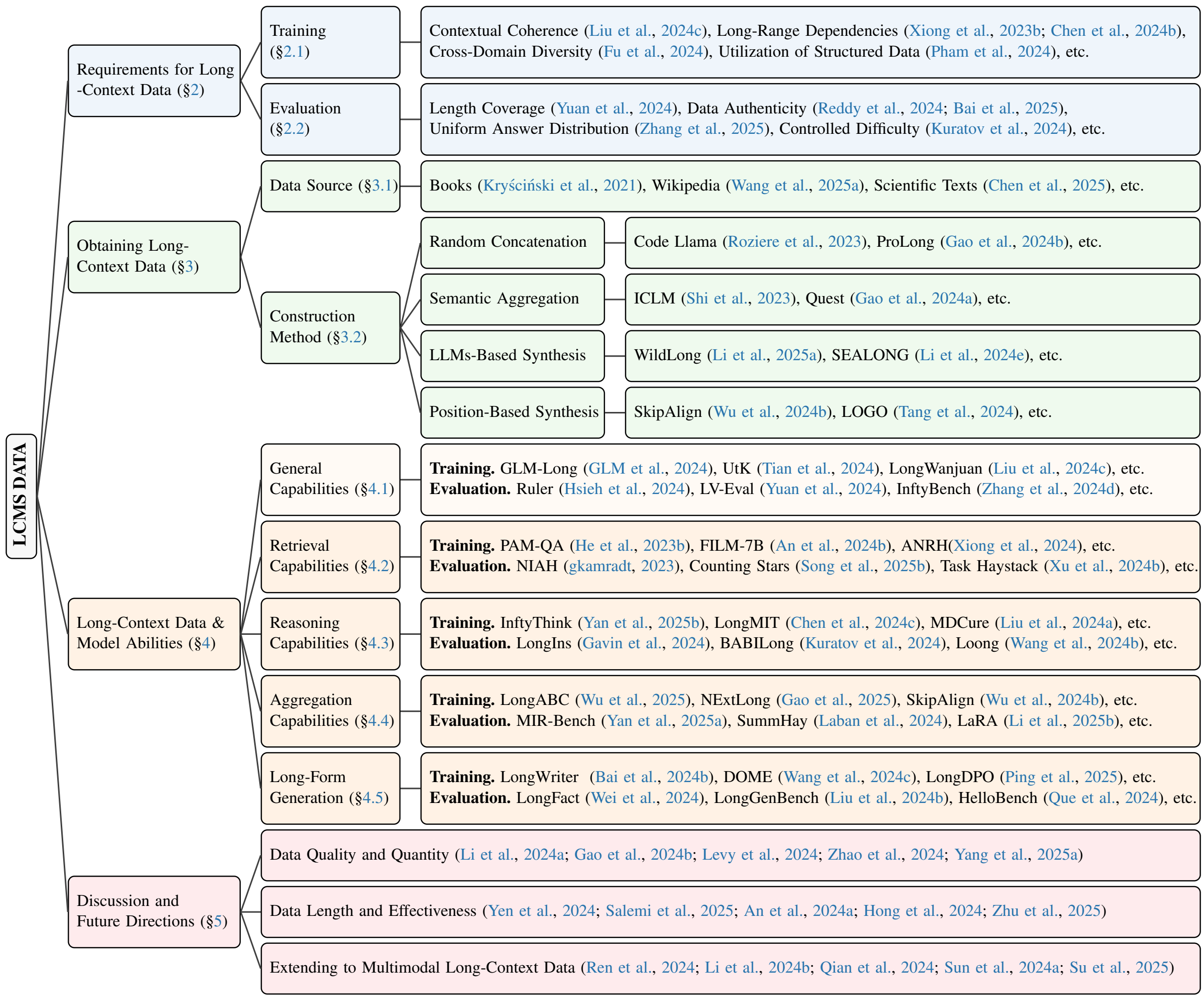
LCMS DATA
Requirements for Long -Context Data (§2)
Training (§2.1)
Contextual Coherence (Liu et al., 2024c), Long-Range Dependencies (Xiong et al., 2023b; Chen et al., 2024b), Cross-Domain Diversity (Fu et al., 2024), Utilization of Structured Data (Pham et al., 2024), etc.
Evaluation (§2.2)
Length Coverage (Yuan et al., 2024), Data Authenticity (Reddy et al., 2024; Bai et al., 2025), Uniform Answer Distribution (Zhang et al., 2025), Controlled Difficulty (Kuratov et al., 2024), etc.
Obtaining Long-Context Data (§3)
Data Source (§3.1)
Books (Kryściński et al., 2021), Wikipedia (Wang et al., 2025a), Scientific Texts (Chen et al., 2025), etc.
Construction Method (§3.2)
Random Concatenation
Code Llama (Roziere et al., 2023), ProLong (Gao et al., 2024b), etc.
Semantic Aggregation
ICLM (Shi et al., 2023), Quest (Gao et al., 2024a), etc.
LLMs-Based Synthesis
WildLong (Li et al., 2025a), SEALONG (Li et al., 2024e), etc.
Position-Based Synthesis
SkipAlign (Wu et al., 2024b), LOGO (Tang et al., 2024), etc.
Long-Context Data & Model Abilities (§4)
General Capabilities (§4.1)
Training. GLM-Long (GLM et al., 2024), UtK (Tian et al., 2024), LongWanjuan (Liu et al., 2024c), etc.
Evaluation. Ruler (Hsieh et al., 2024), LV-Eval (Yuan et al., 2024), InftyBench (Zhang et al., 2024d), etc.
Retrieval Capabilities (§4.2)
Training. PAM-QA (He et al., 2023b), FILM-7B (An et al., 2024b), ANRH(Xiong et al., 2024), etc.
Evaluation. NIAH (gkamradt, 2023), Counting Stars (Song et al., 2025b), Task Haystack (Xu et al., 2024b), etc.
Reasoning Capabilities (§4.3)
Training. InftyThink (Yan et al., 2025b), LongMIT (Chen et al., 2024c), MDCure (Liu et al., 2024a), etc.
Evaluation. LongIns (Gavin et al., 2024), BABILong (Kuratov et al., 2024), Loong (Wang et al., 2024b), etc.
Aggregation Capabilities (§4.4)
Training. LongABC (Wu et al., 2025), NExtLong (Gao et al., 2025), SkipAlign (Wu et al., 2024b), etc.
Evaluation. MIR-Bench (Yan et al., 2025a), SummHay (Laban et al., 2024), LaRA (Li et al., 2025b), etc.
Long-Form Generation (§4.5)
Training. LongWriter (Bai et al., 2024b), DOME (Wang et al., 2024c), LongDPO (Ping et al., 2025), etc.
Evaluation. LongFact (Wei et al., 2024), LongGenBench (Liu et al., 2024b), HelloBench (Que et al., 2024), etc.
Discussion and Future Directions (§5)
Data Quality and Quantity (Li et al., 2024a; Gao et al., 2024b; Levy et al., 2024; Zhao et al., 2024; Yang et al., 2025a)
Data Length and Effectiveness (Yen et al., 2024; Salemi et al., 2025; An et al., 2024a; Hong et al., 2024; Zhu et al., 2025)
Extending to Multimodal Long-Context Data (Ren et al., 2024; Li et al., 2024b; Qian et al., 2024; Sun et al., 2024a; Su et al., 2025)

Figure 1: Taxonomy of Long-Context Data.

model's performance in complex scenarios and promote LCMs optimization.

3 Obtaining Long-Context Data

Compared with short-context data, long-context data is typically more challenging to acquire due to its stringent requirements for length and quality (Ouyang et al., 2022). The acquisition of such data can be categorized into two main approaches: sampling long-context data form existing data sources (Le Scao et al., 2023; Touvron et al., 2023) or synthesizing long-context through strategies (Shi et al., 2023; Tworkowski et al., 2024; Song et al., 2025a). Below, we provide a detailed discussion of long-context data sources (§3.1) and construction methods (§3.2).

3.1 Long-Context Data Source

The sources of long-text data can be categorized into general-purpose data and domain-specific data. General-purpose data primarily originates from sources such as books and lengthy dialogues (Zhao et al., 2023), offering advantages such as large scale, high diversity, and ease of access. These qualities make it highly suitable for LCMs to enhance their language modeling and generalization capabilities. Domain-specific data includes code, academic papers, and specialized knowledge resources, which can significantly improve the performance of LCMs on specific tasks (Yan et al., 2025b; Adams et al., 2024).

Books Books represent one of the most important and frequently used sources of naturally generated long-context data (Kim et al., 2024). They encompass a wide range of styles and genres, providing complete sentences and paragraphs with strong coherence and complex contextual relationships (Bai et al., 2024a; Zhang et al., 2024e). These characteristics enable LCMs to learn deep connections between contexts, thereby improving their ability to understand complex sentence structures, logi-

Datasets	Source	Abilities	Construction method
Pre-Training Data			
SemDeDup (Abbas et al., 2023)	O	General	Sampling
WanJuan (He et al., 2023a)	C, O	General	Sampling
LLaMA2 Long (Xiong et al., 2023b)	B, W, C	Reasoning	Up-Sampling, Generation
ICLM (Shi et al., 2023)	B, W, S	Reasoning, Retrieval	Splicing
UtK (Tian et al., 2024)	B, W, S, C	General	Splicing
Quest (Gao et al., 2024a)	O	General	Splicing
ProLong (Chen et al., 2024b)	B, C, S	General, Reasoning	Up-Sampling
MAP-Neo (Zhang et al., 2024a)	B, S, C, O	General, Reasoning	Sampling
LongWanjuan (Liu et al., 2024c)	B, W, S, C, O	General, Reasoning	Up-Sampling
ProLong 8B (Gao et al., 2024b)	B, W, C	Reasoning	Mixing, Sampling
LLaMA2 128K (Fu et al., 2024)	B, W, C	Reasoning, Retrieval	Splicing, Sampling
RegMIX (Liu et al., 2025b)	B, W, S, C, O	General, Reasoning	Mixing
SPLiCe (Staniszewski et al., 2025)	W, S, C	General,Reasoning	Splicing
NextLong (Gao et al., 2025)	B, W, S	General, Retrieval	Splicing, Positional
LADM (Chen et al., 2025)	B, W, S, C, O	Reasoning, Aggregation	Generation
LongABC (Wu et al., 2025)	B, C, S	Reasoning, Aggregation	Positional
MMR&FPS (Wang et al., 2025b)	O	Reasoning, Aggregation	Splicing, Generation
Post-Training Data			
ASM QA (He et al., 2023b)	O	Reasoning, Aggregation	Generation
LongAlign (Bai et al., 2024a)	B, W, S, C	General, Reasoning	Generation
ChatQA 2 (Xu et al., 2024a)	B, W	General, Reasoning	Sampling, Generation
LongReward (Zhang et al., 2024c)	S	General, Aggregation	Splicing, Generation
LongMIT (Chen et al., 2024c)	B, W, S	Reasoning	Generation
Suri (Pham et al., 2024)	O	Reasoning	Sampling, Annotation
USDC (Marreddy et al., 2024)	O	Reasoning	Generation
MDCure (Liu et al., 2024a)	O	Reasoning, Aggregation	Generation
SkipAlign (Wu et al., 2024b)	S	Reasoning, Aggregation	Positional
IN2 (An et al., 2024b)	O	Reasoning, Retrieval	Generation
SEALONG (Li et al., 2024e)	W	Reasoning, Retrieval	Generation
LOGO (Tang et al., 2024)	B, W, S	Reasoning, Aggregation	Splicing, Positional
ORPO (Hong et al., 2024)	O	Reasoning, Aggregation	Generation
DOME (Wang et al., 2024c)	B, O	Long-Form Generation	Generation
LongWriter (Bai et al., 2024b)	O	Long-Form Generation	Generation
MegaBeam (Wu and Song, 2025)	B, S, C	General, Retrieval	Splicing, Positional
HIERARCHICAL (He et al., 2025)	B	General, Reasoning	Gneration, Positional
LongFaith (Yang et al., 2025a)	W	Reasoning	Generation
InftyThink (Yan et al., 2025b)	O	Reasoning	Generation
WildLong (Li et al., 2025a)	S, C	Reasoning, Aggregation	Generation
DeFine (Wang et al., 2025a)	W	Reasoning, Aggregation	Generation
Light-R1 (Wen et al., 2025)	O	Reasoning	Generation
LongDPO (Ping et al., 2025)	O	Long-Form Generation	Generation
Evaluation Data			
L-EVAL (An et al., 2023)	B, O	Reasoning, Aggregation	Annotation
LongBench v1&v2 (Bai et al., 2023, 2025)	B, W, S	Reasoning, Aggregation	Annotation, Generation
ZeroSCROLLS (Shaham et al., 2023)	B, W, S	Reasoning, Aggregation	Annotation, Generation
LongIns (Gavin et al., 2024)	O	General, Reasoning	Generation
HELMET (Yen et al., 2024)	B, W, O	General, Reasoning	Annotation
InftyBench (Zhang et al., 2024d)	B, C	General, Retrieval	Annotation, Generation
FanOutQA (Zhu et al., 2024)	W	Reasoning, Retrieval	Annotation
NovelQA (Wang et al., 2024a)	B	Reasoning, Retrieval	Annotation
BABILong (Kuratov et al., 2024)	B	Reasong, Retrieval	Annotation, Generation
LV-EVAL (Yuan et al., 2024)	B, W, O	Reasoning, Retrieval	Annotation, Genreation
HoloBench (Maekawa et al., 2024)	W, O	Reasoning, Retrieval	Generation
FinTextQA (Chen et al., 2024a)	O	Reasoning, Aggregation	Annotation
Loong (Wang et al., 2024b)	S, O	Reasoning, Aggregation	Annotation
NarrativeQA (Bohnet et al., 2024)	B	Reasoning, Aggregation	Generation
Ruler (Hsieh et al., 2024)	O	Reasoning, Aggregation	Generation
Task Haystack (Li et al., 2024d)	W, O	Retrieval	Annotation
SummHay (Laban et al., 2024)	O	Aggregation, Retrieval	Generation
XL2Bench (Ni et al., 2024)	B, S, O	Reasoning, Retrieval	Generation
Michelangelo (Vodrahalli et al., 2024)	W, C	Reasoning, Aggregation	Generation
LCFO (Costa-jussà et al., 2024)	W, S, O	Aggregation	Annotation
HelloBench (Que et al., 2024)	B, W, S, O	Long-Form Generation	Annotation
LongGenBENCH (Liu et al., 2024b)	O	Long-Form Generation	Generation
LongFact (Wei et al., 2024)	O	Long-Form Generation	Generation
LongLaMP (Kumar et al., 2024)	S, O	Long-Form Generation	Generation
LongCodeBench (Rando et al., 2025)	C	Reasoning	Annotation, Generation
LaRA (Li et al., 2025b)	B, S, O	Reasoning, Retrieval	Annotation, Generation
DeFine (Wang et al., 2025a)	O	Reasoning, Aggregation	Annotation
MIR-Bench (Yan et al., 2025a)	C	Reasoning, Aggregation	Generation
Needle Threading (Roberts et al., 2025)	B	Retrieval	Generation
LONGINOUTBENCH (Zhang et al., 2025)	S	Long-Form Generation	Generation

Table 1: Datasets and Benchmarks for LCMs Training and Evaluation. “B”, “W”, “S”, “C”, and “O” refer to Books, Wikipedia, Scientific-Texts, Code, and Other domain-specific data (e.g. medical, financial, etc.).

cal relationships, and semantic coherence. Currently, widely used book datasets include Books3 and BookCorpus2 [1].

Wikipedia Wikipedia [2] serves as a high-quality knowledge base and is extensively utilized in the training and evaluation of LCMs (Xiong et al., 2023a; Gao et al., 2024b; Wang et al., 2025a). Its entries cover a broad spectrum of topics, enhancing the model's knowledgeability and logical reasoning abilities. Additionally, Wikipedia's hierarchical structure makes it particularly suitable for constructing multi-hop reasoning datasets with long-context dependencies (Zhu et al., 2024).

Scientific Texts Scientific text encompasses textbooks, academic papers, and related resources. Such data plays a crucial role in training and testing LCMs' ability to comprehend scientific knowledge (Tian et al., 2024; Chen et al., 2025). Common sources of scientific texts include arXiv [3] papers, PubMed [4] papers, textbooks, lecture notes, and educational webpages.

Code Code also meets the length requirements of long-context data and differs significantly from natural language (Roziere et al., 2023). As a formalized language, code relies on strict syntax and specific programming paradigms, reflecting long-range dependencies and precise execution logic. Primary sources of code include programming Q&A communities (e.g., Stack Exchange [5]) and public software repositories (e.g., GitHub [6]). The former provides rich context and real-world usage scenarios, while the latter covers multiple programming languages and domains, ensuring high quality and diversity (Wu and Song, 2025; Wu et al., 2025).

3.2 Long-Context Data Construction Method

Existing native long-context data is not only scarce but also costly to acquire. Additionally, processing such data poses significant technical challenges, including difficulties in annotation, sparsity of key information, and the complexity of maintaining logical coherence (Zhang et al., 2024a; Quan et al., 2024). Consequently, synthesizing long-context data through specific strategies—such as document concatenation or leveraging large language models (LLMs)—has become a critical approach to overcoming the bottleneck of long-context data availability and enhancing the performance of LCMs (Pham et al., 2024; Liang et al., 2024).

Random Concatenation Strategy The random concatenation strategy involves combining short documents randomly to achieve a target length, enabling rapid generation of long-context data (Ouyang et al., 2022; Le Scao et al., 2023; Touvron et al., 2023). While this method ensures diversity in the contextual content of the synthesized data, the weak semantic relationships between concatenated documents hinder the model's ability to learn long-range dependencies (Levine et al., 2021). As such, data generated via random concatenation is typically suitable for the pre-training phase of LCMs, which requires large amounts of unsupervised data and does not involve manual annotation.

Semantic Aggregation Strategy The semantic aggregation strategy generates long-context data by aggregating semantically similar documents (Shi et al., 2023). For instance, a document can be concatenated with the top k most similar documents in the corpus (Guu et al., 2020; Yang et al., 2024a). This approach emphasizes semantic relevance, enhancing the coherence of the synthesized long text. However, excessive reliance on semantic similarity may lead to narrow contexts (i.e., high redundancy), thereby compromising the diversity of the long-context data (Gao et al., 2024a).

LLMs-Based Synthesis Strategy Owing to the rapid advancements in LLMs, LLM-based long-context synthesis has emerged as an efficient method to significantly reduce manual annotation costs while enriching long-context data resources. Data synthesized using LLMs is typically supervised, including instruction-tuning data (An et al., 2024b; He et al., 2023a) and preference-alignment data (Zhang et al., 2024c; Ping et al., 2025; Hong et al., 2024). The most common form of such data is question-answer pairs (QA pairs), which are used during the post-training phase. In this phase, pre-trained language models with general capabilities undergo targeted training to acquire domain-specific skills or produce outputs that better align with human preferences (Yang et al., 2024b).

Position-Based Synthesis Strategy By inserting specific position indices during training, models

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

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

[3] https://arxiv.org/

[4] https://pubmed.ncbi.nlm.nih.gov/

[5] https://stackexchange.com/

[6] https://github.com/

can handle longer contexts without increasing additional computational resources (Zhu et al., 2023; Su et al., 2024; Ding et al., 2024). This methods construct "long-context" data by manipulating position indices, rather than relying on extending the actual length of input sequences (Wu et al., 2024a). For example, SkipAlign (Wu et al., 2024b) strategically inserts skipped positions into instruction-following samples, leveraging the semantic structure of the data to effectively extend context and synthesize long-range dependencies. DAPE (Zheng et al., 2024) achieves length extrapolation through data-adaptive position encoding.

4 Long-Context Data Driven Model Capabilities

LCMs are capable of effectively processing text sequences containing thousands or even tens of thousands of tokens, thereby excelling in more complex language understanding and generation tasks (anthropic, 2024; OpenAI, 2024). This capability also imposes higher demands on various aspects of the model's performance. This chapter systematically examines the impact of data on the development and evaluation of LCMs capabilities from the perspective of model competencies. Specifically, we categorize the core capabilities of these models into five types: General Capabilities (§4.1), which are fundamental language modeling abilities acquired during the pre-training phase and serve as the foundation for other advanced abilities; Retrieval Capabilities (§4.2), Reasoning Capabilities (§4.3), Aggregation Capabilities (§4.4), and Long-Form Generation Capabilities (§4.5), as shown in Figure 2. Building on this framework, we conduct an analysis from two dimensions—training data and evaluation data—focusing on the following key questions 1) **What training data are most conducive to enhancing LCMs performance** and 2) **What benchmarks are more accurate and comprehensive** of these capabilities. Through this systematic analytical framework, we aim to provide both theoretical support and practical guidance for optimizing the capabilities of LCMs.

4.1 General Capabilities

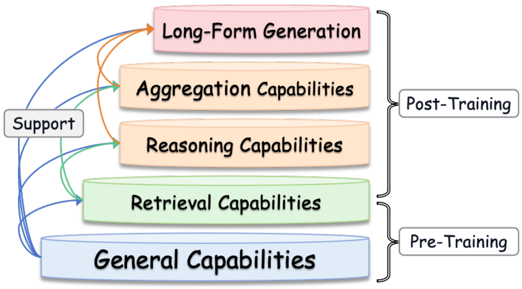

Figure 2: LCMs Capabilities and Training Phases

General capabilities of LCMs refer to the foundational language understanding and generation abilities that possess in long-context scenarios. These include mastery of basic grammatical structures, comprehension of semantic relationships, and commonsense reasoning (Beltagy et al., 2020; MiniMax et al., 2025). General capabilities enable LCMs to accurately process and generate text that conforms to linguistic norms, while serving as the foundation for supporting other advanced capabilities.

General capabilities are primarily acquired during the pre-training phase of LLMs through large-scale unsupervised learning. Typically, diverse long-context data are obtained from existing large corpora using splicing and sampling methods. These data sources are extensive, covering various types such as books, Wikipedia, and code. For example, Qwen-2.5 (Yang et al., 2024a) and GLM-Long (GLM et al., 2024) enhance the models' ability to process million-token-level long-contexts by splicing and upsampling natural long-context data, while introducing synthetic data. Research efforts such as UtK (Tian et al., 2024), Quest (Gao et al., 2024a), and SPLiCe (Staniszewski et al., 2025) further emphasize the importance of long-range dependencies in long-context data. These methods split long contexts into shorter segments and reassemble them to generate training samples with higher diversity. Additionally, Fu et al. (2024) and Ye et al. (2025) explore the significant impact of mixing proportions of data from different domains. Datasets such as LongWanJuan (Liu et al., 2024c) and RegMix (Liu et al., 2025b) construct high-quality, domain-balanced, large-scale training sets for LCMs from the perspective of data mixing, providing robust data support for the development of general capabilities and significantly enhancing the overall performance of LCMs.

The evaluation of LCMs general capabilities relies on multi-task comprehensive benchmarks that integrate diverse tasks and datasets to offer a holistic assessment. For example, ZeroSCROLLS (Shaham et al., 2023) extends SCROLLS (Shaham et al., 2022) by including 10 natural language tasks such as QA and information aggregation, focusing on zero-shot understanding. BABILong (Kuratov et al., 2024) emphasizes complex reasoning and retrieval through 20 reasoning tasks, including multi-hop QA and fact chain reasoning. RULER (Hsieh et al., 2024) builds on the NIAH (gkamradt, 2023) benchmark to evaluate ultra-long context search abilities. InftyBench (Zhang et al., 2024d), covering multiple domains and languages, assesses retrieval and reasoning performance on texts exceeding 100k tokens. Other benchmarks combine existing datasets for broader evaluations. LV-Eval (Yuan et al., 2024) evaluates performance across five text lengths. Longbench (Bai et al., 2023) and its updated version Longbench-v2 (Bai et al., 2025) focus on bilingual real-world scenarios with more realistic long-context tasks. L-Eval (An et al., 2023) covers law, finance, and other domains through 20 subtasks. HELMET (Yen et al., 2024) expands into seven application-driven categories and supports input lengths up to 128k tokens.

Summary. The development of general capabilities in LCMs require sufficiently long, diverse, and domain-balanced datasets, as well as multi-task evaluation benchmarks that are flexible in length, moderate in task difficulty, and broadly applicable.

4.2 Retrieval Capabilities

Retrieval capability of LCMs refers to the ability to quickly locate information relevant to user queries in a massive context, and integrate it (Fei et al., 2024; Roberts et al., 2025). This process not only relies on surface-level keyword matching but also demands that the model possess robust "attention mechanisms" and "information filtering" capabilities. These enable precise and comprehensive retrieval of long-context, thereby identifying semantic segments most relevant to the key information (Xu et al., 2023; Goldman et al., 2024).

Existing work typically leverages synthetic data to enhance the retrieval capabilities of LCMs during the post-training phase (Xu et al., 2024a). A key challenge in this process is the "Lost in the Middle" phenomenon, where models tend to overly focus on information at the beginning and end of ultra-long texts while neglecting critical information located in the middle (Liu et al., 2023; Zhang et al., 2025). To address this, PAM-QA (He et al., 2023b) improves retrieval performance by decomposing multi-document QA tasks into multiple reasoning steps, including question paraphrasing, index prediction, and answer summarization, thereby guiding the model to better focus on target information. An et al. (2024b) synthesized long-context QA data that explicitly teaches models that key information can appear at any position within the context. The resulting model, FILM-7B, demonstrates robust retrieval of information from any position within the context window. Existing research has shown that structured training data plays a positive role in enhancing the retrieval capabilities of LCMs (Li et al., 2024d). For instance, Baker et al. (2024) utilize triple extraction and summarization techniques from knowledge graphs to construct concise and well-structured training samples, to guide models to perform more precise information retrieval. Xiong et al. (2024) design a dataset based on key-value retrieval tasks, which not only improves the retrieval accuracy of models but also effectively mitigates hallucination.

Retrieval benchmarks aim to assess LCMs ability to locate and extract key information within long-contexts (Askari et al., 2024; Yuan et al., 2024), with the "Needle In A Haystack (NIAH)" (gkamradt, 2023) being one of the most representative. NIAH requires LCMs to retrieve critical information ("needle") from ultra-long irrelevant text ("haystack"), thereby evaluating its retrieval performance across varying context lengths and depths. Building on this, NeedleBench (Li et al., 2024c) introduces multi-needle retrieval tasks (M-RT) and multi-needle reasoning tasks (M-RS) to further explore the complex retrieval and certain reasoning capabilities of LCMs. Meanwhile, RULER (Hsieh et al., 2024) incorporates four variants of the NIAH task to deeply investigate LCMs performance under different retrieval conditions. Counting-Stars (Song et al., 2025b) requires models to accurately retrieve and output the number of inserted "stars" in long-contexts, offering a more precise measure of LCMs ability to handle long-range dependencies compared to traditional NIAH, thus providing new insights into their potential for complex information processing and detailed task execution. Furthermore, some comprehensive evaluation benchmarks also include extensive retrieval tasks (Kuratov et al., 2024; Bai et al., 2023), such as the Retrieve.PassKey (Mohtashami and Jaggi, 2023) in Infty-Bench (Zhang et al., 2024d) and TriviaQA (Joshi et al., 2017) in HELMET (Yen et al., 2024). Many long-context reasoning benchmarks (Li et al., 2025b; Ni et al., 2024; Laban et al., 2024), including FanoutQA (Zhu et al., 2024) and

Loong (Wang et al., 2024b), require models to first retrieve key information from multiple documents before assessing their comprehension abilities.

Summary. Synthetic data with explicit key information and clear structure is more conducive to enabling LCMs to achieve efficient and accurate information retrieval in complex long-context scenarios (Qu et al., 2025). To properly evaluate this retrieval capability, ultra-long evaluation data with flexible and controllable key information is required, which helps avoid the phenomenon of shortcut learning (Du et al., 2023; Sun et al., 2024b) based on the original knowledge.

4.3 Reasoning Capabilities

Reasoning capability of LCMs refers perform logical deduction and judgment based on given contextual information and internalized knowledge (Wan et al., 2025a; Yang et al., 2025b). This enables LCMs to understand complex logical structures, such as causal and conditional relationships, and to derive new conclusions or insights. Such reasoning ability allows LCMs to effectively address tasks that require in-depth thinking and logical analysis.

LCMs exhibit substantially weaker reasoning capabilities on long-context tasks compared to their performance on short-context tasks (GLM et al., 2024; Liu et al., 2024d). To bridge this gap, recent research commonly leverage advanced LLMs to synthesize high-quality instruction data, which is used for post-training to enhance the complex reasoning abilities of LCMs (Zhang et al., 2024e; Tang et al., 2024). InftyThink (Yan et al., 2025b) reconstructs long-context reasoning datasets into an iterative format, facilitating better learning of complex reasoning paths. WildLong extracts meta-information from real user queries, models co-occurrence relationships using graph-based methods, to construct scalable training data. LongFaith (Yang et al., 2025a) improves the accuracy of reasoning chains by integrating ground-truth with citation-based reasoning prompts. Moreover, both LongMIT (Chen et al., 2024c) and MDCure (Liu et al., 2024a) generate high-quality instructions based on multi-document QA tasks. Among these, multi-hop instructions are particularly beneficial for enhancing the reasoning capabilities of LCMs. Beyond instruction tuning, preference optimization data also enhance the complex reasoning capabilities of LCMs (Zhang et al., 2024c; Marreddy et al., 2024). SeaLong (Li et al., 2024e) generates multiple responses per question, ranks them and then conducts either supervised fine-tuning or preference optimization based on the ranked outputs. Light-R1 (Wen et al., 2025) proposes a curriculum learning approach, demonstrating that using unique and diverse datasets at different training stages can be more effective.

Reasoning benchmarks are designed to assess capacity for logical inference and deep comprehension in long-context contexts, going beyond simple retrieval of key information (Ho et al., 2020; Amos et al., 2023). Among these, single-document reasoning tasks are the most commonly used (Dasigi et al., 2021). However, due to the concentrated distribution of ground truth within long-contexts, these tasks often provide a limited assessment of LCMs' reasoning capabilities (Trivedi et al., 2022; Zhang et al., 2024b). To address this limitation, recent benchmarks introduce more diverse and challenging reasoning structures (Ni et al., 2024; Wang et al., 2024a). For example, FanoutQA (Zhu et al., 2024) leverages the hierarchical structure of Wikipedia to construct multi-hop reasoning tasks, Michelangelo (Vodrahalli et al., 2024) challenges models to extract relevant information from large volumes of irrelevant content, BABILong (Kuratov et al., 2024) encompasses 20 distinct reasoning tasks, including chain-of-facts reasoning and basic induction, aiming to evaluate cross-fact reasoning abilities. Furthermore, Loong (Wang et al., 2024b) and LongBench (Bai et al., 2023, 2025) focus on evaluating deep understanding and reasoning capabilities across multiple real-world tasks. These benchmarks emphasize the importance of handling realistic scenarios with extended contexts. Given the practical importance of reasoning in real-world applications, some efforts have been directed toward evaluating LCMs in domain-specific settings. LongCodeBench is built on GitHub issues and code repositories, testing abilities in code comprehension and bug fixing within million-token contexts. DocFinQA (Reddy et al., 2024) and FinTextQA (Chen et al., 2024a) are financial-domain QA datasets on long-document comprehension in realistic financial scenarios. LongHealth (Adams et al., 2024) provides a comprehensive evaluation to process lengthy clinical documents encountered in healthcare settings.

Summary. Multi-document and multi-hop QA with explicit intermediate reasoning processes are highly beneficial for enhancing the reasoning capabilities of LCMs. Reasoning chains and knowledge graphs further facilitates the explicit representation of reasoning paths, thereby improving both the interpretability and accuracy of reasoning. To effectively evaluate these capabilities, it is essential to design datasets with increased complexity and robustness, paired with tasks that demand advanced reasoning skills.

4.4 Aggregation Capabilities

The aggregation capability of LCMs refers to effectively integrating and synthesizing distributed information into a more comprehensive and well-structured knowledge representation (Li et al., 2024a; Zhang et al., 2024f; Liu et al., 2025a). This crucially involves categorizing similar content, summarizing key points, and eliminating redundancy (Li et al., 2025b).

Notably, compared to reasoning capabilities, aggregation relies more heavily on accurately capturing long-range dependencies within the training data. For instance, LongABC (Wu et al., 2025) leverages the self-attention mechanisms of LCMs to effectively quantify these long-range dependencies, thereby demonstrably enhancing aggregation performance. Similarly, the NExtLong approach (Gao et al., 2025) introduces negative document extension as a strategy to synthesize long-context data. This forces LCMs to meticulously distinguish relevant long-range context from irrelevant content, thus significantly strengthening their dependency modeling capabilities. Moreover, innovative positional-based data synthesis methods allow for precise control over the uniform distribution of key information within long contexts, which has also proven particularly beneficial for improving aggregation performance (Chen et al., 2023). Furthermore, SkipAlign (Wu et al., 2024b) enables LCMs to better capture complex long-range dependencies by astutely leveraging both positional indexing and underlying semantic structure. Additionally, it is worth noting that QA datasets presented in summarization task formats are also highly beneficial for training and refining the aggregation capabilities of LCMs (Kryściński et al., 2021; Zhang et al., 2024e).

Aggregation capability benchmarks are designed to assess LLMs to integrate, summarize, and perform cross-document analysis across multiple long-contexts (Koh et al., 2022; Yan et al., 2025a). These tasks typically involve summarization and synthesis, and often impose demands on the retrieval and reasoning capabilities (Shaham et al., 2023). GovReport (Huang et al., 2021) and QMSum (Zhong et al., 2021) , constructed from government reports and meeting transcripts, respectively, evaluate LCMs ability to aggregate information in real-world scenarios. LCFO (Costa-jussà et al., 2024) evaluates progressive summarization and summary expansion across diverse domains. HoloBench (Maekawa et al., 2024) introduces database-style reasoning operations into textual contexts, facilitating a systematic evaluation of how LCMs handle holistic inference and aggregation across large document collections. SummHay (Laban et al., 2024) leverages synthetic multi-document "haystack" data to test LCMs to generate accurate summaries from extensive long-document corpora, with the additional requirement to identify relevant insights and properly cite source documents. XL2Bench (Ni et al., 2024) provides a comprehensive evaluation of LCMs aggregation capabilities across three distinct domains—focusing on long-range dependencies modeling.

Summary. The development of aggregation ability depends on effectively modeling long-range dependencies in long-context data, with position-index synthesis strategies demonstrating effectiveness in simulating such dependencies. These evaluation primarily involves summarization and synthesis tasks over long-contexts, requiring realistic data for effective benchmarking. A key challenge lies in assessing the quality of generated summaries, which critically depends on the reliability and accuracy of ground-truth references.

4.5 Long-Form Generation Capabilities

Long-form generation capability of LCMs refers to to produce coherent, fluent, and logically structured long-context content, such as articles and reports. This capacity enables LCMs to meet user demands for text creation of substantial length and depth, while maintaining global structural integrity, narrative flow, and factual consistency across extended contexts. (Li et al., 2023; Wan et al., 2025b)

Long-form generation poses a particularly challenging capability for LCMs. Existing datasets often lack the hierarchical structure and fine-grained annotations necessary for effective task decomposition, resulting in generated texts that are superficial and disorganized (Wang et al., 2024c). To address this limitation, DeFine (Wang et al., 2025a) introduces a decomposed, fine-grained annotation dataset for long-article generation. It incorporates a hierarchical decomposition strategy combined with domain-specific knowledge and multi-level annotations, enabling precise control over generation granularity and promoting content depth. In addition, LongWriter (Bai et al., 2024b) proposes an AgentWrite pipeline that first generates a paragraph-level outline and then produces each section sequentially, while LongDPO (Ping et al., 2025) integrates a global memory pool and external critique mechanisms to generate higher-quality long texts, leading to substantial improvements in long-form generation tasks.

For evaluation purposes, the LongLaMP benchmark (Kumar et al., 2024) defines personalized long-form generation tasks by meticulously combining diverse prompts with retrieval-augmented frameworks. This rigorously assesses the model's ability to integrate retrieved information into coherent, extended outputs. In a similar vein, LongFact (Wei et al., 2024) introduces a novel set of GPT-4-generated prompts that specifically require multi-paragraph, fact-rich responses. It also employs an automated SAFE evaluator to quantitatively measure factual accuracy in such long-form generation (Bai et al., 2024b). Furthermore, the LongGenBench suite (Liu et al., 2024b) presents diverse generation tasks, including technical expositions and creative stories, primarily to evaluate logical consistency maintained over considerably extended contexts. Separately, the HelloBench framework (Que et al., 2024) curates its prompts from authentic sources like books, reports, and transcripts. It then applies a detailed, layered human evaluation framework to assess critical aspects such as narrative flow, overall coherence, and factual coverage. Finally, to test resilience, LongInOutBench (Zhang et al., 2025) introduces controlled perturbations within mid-document content, such as shuffled paragraphs. This specifically tests the model's robustness and its ability to maintain contextual fidelity despite such disruptions.

Summary. Training data that combine hierarchical outlines and stepwise preferences from rich multi-document contexts best develop long-form generation capabilities, while evaluation datasets that holistically cover personalization, factuality, coherence, and robustness comprehensively measure whether LCMs can sustain well-structured, accurate output over thousands of tokens.

5 Discussion and Future Directions

In this work, we present a systematic and comprehensive review of existing research on data-related aspects for long-context large language models (LCMs), with a particular emphasis on the relationship between data characteristics and model capabilities. We discuss datasets that are specifically designed to train and evaluate different abilities, offering clear guidelines for data selection and design in LCMs while laying a theoretical foundation for further enhancing their performance. Nevertheless, many challenges remain in the realm of data for LCMs. This section highlights some of key challenges and discusses future research directions.

Data Quality and Quantity The scarcity of genuinely high-quality, extensive long-context data represents a significant and persistent challenge (Zhao et al., 2024; Li et al., 2024a). While numerous studies have utilized large language models to synthesize long-context data, the verifiable quality of such synthetic data often remains inconsistent. Currently, there is a notable lack of clear definitions and robust evaluation metrics for long-context datasets (Gao et al., 2024b). Future work should therefore critically aim to develop a unified framework for quantifying the quality of long-context datasets across multiple dimensions.

Data Length and Effectiveness Increasing input length significantly raises computational costs during training (Zhang et al., 2024a; An et al., 2024a). Consequently, researchers are exploring the use of hybrid datasets that combine short- and long-context data. However, relying on limited amounts of long-context data may not be sufficient to ensure effective generalization from short to long contexts (Salemi et al., 2025; Zhu et al., 2025). Therefore, it is essential to strike a balance between data length, the proportion of short-to-long mixtures, and overall training effectiveness.

Extending to Multimodal Long-Context Data Future research can address the complexities of multimodal long-context data, encompassing text, image, and video. Key challenges include curating aligned datasets, modeling inter-modal long-range dependencies, synthesizing realistic multimodal sequences, and developing robust evaluation metrics for these integrated contexts. This will unlock more sophisticated, human-like understanding in LCMs (Sun et al., 2024a; Su et al., 2025).

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. 2024. Longhealth: A question answering benchmark with long clinical documents. *arXiv preprint arXiv:2401.14490*.

Ido Amos, Jonathan Berant, and Ankit Gupta. 2023. Never train from scratch: Fair comparison of long-sequence models requires data-driven priors. *arXiv preprint arXiv:2310.02980*.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*.

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

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024b. Make your llm fully utilize the context. *arXiv preprint arXiv:2404.16811*.

Anthropic. The claude 3 model family: Opus, sonnet, haiku.

anthropic. 2024. Claude-3-5-sonnet model card. *blog*.

Arian Askari, Suzan Verberne, Amin Abolghasemi, Wessel Kraaij, and Gabriella Pasi. 2024. Retrieval for extremely long queries and documents with rprs: a highly efficient and effective transformer-based re-ranker. *ACM Transactions on Information Systems*, 42(5):1–32.

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. *arXiv preprint arXiv:2401.18058*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

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. 2025. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*.

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.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

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*.

Valeriia Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. 2023. Wikihowqa: A comprehensive benchmark for multi-document non-factoid question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5314.

Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024a. Fintextqa: A dataset for long-form financial question answering. *arXiv preprint arXiv:2405.09980*.

Jianghao Chen, Junhong Wu, Yangyifan Xu, and Jiajun Zhang. 2025. Ladm: Long-context training data selection with attention-based dependency measurement for llms.

Longze Chen, Ziqiang Liu, Wanwei He, 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. *arXiv preprint arXiv:2405.17915*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Zhi Chen, Qiguang Chen, Libo Qin, Qipeng Guo, Haijun Lv, Yicheng Zou, Wanxiang Che, Hang Yan, Kai Chen, and Dahua Lin. 2024c. What are the essential factors in crafting effective long context multi-hop instruction datasets? insights and best practices. *arXiv preprint arXiv:2409.01893*.

Marta R Costa-jussà, Pierre Andrews, Mariano Coria Meglioli, Joy Chen, Joe Chuang, David Dale, Christophe Ropers, Alexandre Mourachko, Eduardo Sánchez, Holger Schwenk, et al. 2024. Lcfo: Long context and long form output dataset and benchmarking. *arXiv preprint arXiv:2412.08268*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding.

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. *arXiv preprint arXiv:2402.10171*.

Chaochen Gao, Xing Wu, Qi Fu, and Songlin Hu. 2024a. Quest: Query-centric data synthesis approach for long-context scaling of large language model. *arXiv preprint arXiv:2405.19846*.

Chaochen Gao, Xing Wu, Zijia Lin, Debing Zhang, and Songlin Hu. 2025. Nextlong: Toward effective long-context training without long documents. *arXiv preprint arXiv:2501.12766*.

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

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*.

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.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. 2023a. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. *arXiv preprint arXiv:2308.10755*.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, Yibo Liu, Qianguo Sun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2023b. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. *arXiv preprint arXiv:2311.09198*.

Linda He, Jue Wang, Maurice Weber, Shang Zhu, Ben Athiwaratkun, and Ce Zhang. 2025. Scaling instruction-tuned llms to million-token contexts via hierarchical synthetic data generation. *arXiv preprint arXiv:2504.12637*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2(4):5.

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? *arXiv preprint arXiv:2404.06654*.

Yutong Hu, Quzhe Huang, Kangcheng Luo, and Yansong Feng. 2024. What kinds of tokens benefit from distant text? an analysis on long context language modeling. *arXiv preprint arXiv:2406.11238*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

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. *arXiv preprint arXiv:2404.01261*.

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. 2021. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

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. *arXiv preprint arXiv:2406.10149*.

Philippe Laban, Alexander R Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and rag systems. *arXiv preprint arXiv:2407.01370*.

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.

Yoav Levine, Noam Wies, Daniel Jannai, Dan Navon, Yedid Hoshen, and Amnon Shashua. 2021. The inductive bias of in-context learning: Rethinking pretraining example design. *arXiv preprint arXiv:2110.04541*.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models.

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?

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. *arXiv preprint arXiv:2502.09977*.

Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. 2024b. Temporal reasoning transfer from text to video.

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

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, Wenbo Su, and Bo Zheng. 2024d. Graphreader: Building graph-based agent to enhance long-context abilities of large language models.

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

Yi Liang, You Wu, Honglei Zhuang, Li Chen, Jiaming Shen, Yiling Jia, Zhen Qin, Sumit Sanghai, Xuanhui Wang, Carl Yang, and Michael Bendersky. 2024. Integrating planning into single-turn long-form text generation.

Gabrielle Kaili-May Liu, Bowen Shi, Avi Caciularu, Idan Szpektor, and Arman Cohan. 2024a. Mdcure: A scalable pipeline for multi-document instruction-following. *arXiv preprint arXiv:2410.23463*.

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. 2025a. A comprehensive survey on long context language modeling.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. 2025b. Regmix: Data mixture as regression for language model pre-training.

Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. 2024b. Longgenbench: Long-context generation benchmark. *arXiv preprint arXiv:2410.04199*.

Xiaoran Liu, Kai Lv, Qipeng Guo, Hang Yan, Conghui He, Xipeng Qiu, and Dahua Lin. 2024c. 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. 2024d. Chatqa: Building gpt-4 level conversational qa models. *CoRR*.

Seiji Maekawa, Hayate Iso, and Nikita Bhutani. 2024. Holistic reasoning with long-context lms:

A benchmark for database operations on massive textual data. *arXiv preprint arXiv:2410.11996*.

Mounika Marreddy, Subba Reddy Oota, Venkata Charan Chinni, Manish Gupta, and Lucie Flek. 2024. Usdc: A dataset of User Stance and Dogmatism in long Conversations. *arXiv preprint arXiv:2406.16833*.

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.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Landmark attention: Random-access infinite context length for transformers. *ArXiv*, abs/2305.16300.

Xuanfan Ni, Hengyi Cai, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, and Piji Li. 2024. XlΘ2 bench: A benchmark for extremely long context understanding with long-range dependencies. *arXiv preprint arXiv:2404.05446*.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

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.

Chau Minh Pham, Simeng Sun, and Mohit Iyyer. 2024. Suri: Multi-constraint instruction following for long-form text generation.

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. *arXiv preprint arXiv:2502.02095*.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. 2024. Momentor: Advancing video large language model with fine-grained temporal reasoning.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, et al. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*.

Shanghaoran Quan, Tianyi Tang, Bowen Yu, An Yang, Dayiheng Liu, Bofei Gao, Jianhong Tu, Yichang Zhang, Jingren Zhou, and Junyang Lin. 2024. Language models can self-lengthen to generate long texts.

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. *arXiv preprint arXiv:2401.06915*.

Weiming Ren, Huan Yang, Jie Min, Cong Wei, and Wenhu Chen. 2024. Vista: Enhancing long-duration and high-resolution video understanding by video spatiotemporal augmentation.

Jonathan Roberts, Kai Han, and Samuel Albanie. 2025. Needle threading: Can llms follow threads through near-million-scale haystacks?

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*.

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

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. 2022. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A Smith, Luke Zettlemoyer, et al. 2023. In-context pretraining: Language modeling beyond document boundaries. *arXiv preprint arXiv:2310.10638*.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025a. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv e-prints*, pages arXiv–2501.

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.

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.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. 2025. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*.

Jiashuo Sun, Jihai Zhang, Yucheng Zhou, Zhaochen Su, Xiaoye Qu, and Yu Cheng. 2024a. Surf: Teaching large vision-language models to selectively utilize retrieved information. *arXiv preprint arXiv:2409.14083*.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024b. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6883–6893, Torino, Italia. ELRA and ICCL.

Zecheng Tang, Zechen Sun, Juntao Li, Qiaoming Zhu, and Min Zhang. 2024. Logo–long context alignment via efficient preference optimization. *arXiv preprint arXiv:2410.18533*.

Junfeng Tian, Da Zheng, Yang Cheng, Rui Wang, Colin Zhang, and Debing Zhang. 2024. Untie the knots: An efficient data augmentation strategy for long-context pre-training in language models. *arXiv preprint arXiv:2409.04774*.

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*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2024. Focused transformer: Contrastive training for context scaling. *Advances in Neural Information Processing Systems*, 36.

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.

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.

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. 2025a. Define: A decomposed and fine-grained annotated dataset for long-form article generation.

Minzheng Wang, Longze Chen, Cheng Fu, 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. *arXiv preprint arXiv:2406.17419*.

Qianyue Wang, Jinwu Hu, Zhengping Li, Yufeng Wang, daiyuan li, Yu Hu, and Mingkui Tan. 2024c. Generating long-form story using dynamic hierarchical outlining with memory-enhancement.

Zhchao Wang, Bin Bi, Yanqi Luo, Sitaram Asur, and Claire Na Cheng. 2025b. 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. *arXiv preprint arXiv:2403.18802*.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, et al. 2025. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *arXiv preprint arXiv:2503.10460*.

Chen Wu and Yin Song. 2025. Scaling context, not parameters: Training a compact 7b language model for efficient long-context processing. *arXiv preprint arXiv:2505.08651*.

Longyun Wu, Dawei Zhu, Guangxiang Zhao, Zhuocheng Yu, Junfeng Ran, Xiangyu Wong, Lin Sun, and Sujian Li. 2025. Longattn: Selecting long-context training data via token-level attention.

Tong Wu, Yanpeng Zhao, and Zilong Zheng. 2024a. Never miss a beat: An efficient recipe for context window extension of large language models with consistent" middle" enhancement. *arXiv e-prints*, pages arXiv–2406.

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

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. 2023a. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. 2023b. Effective long-context scaling of foundation models, 2023. *URL https://arxiv. org/abs/2309.16039*.

Zheyang Xiong, Vasilis Papageorgiou, Kangwook Lee, and Dimitris Papailiopoulos. 2024. From artificial needles to real haystacks: Improving retrieval capabilities in llms by finetuning on synthetic data.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2024a. Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities. *arXiv preprint arXiv:2407.14482*.

Xiaoyue Xu, Qinyuan Ye, and Xiang Ren. 2024b. Stress-testing long-context language models with lifelong icl and task haystack.

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, 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. 2025a. Longfaith: Enhancing long-context reasoning in llms with faithful synthetic data. *arXiv preprint arXiv:2502.12583*.

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.

Wang Yang, Zirui Liu, Hongye Jin, Qingyu Yin, Vipin Chaudhary, and Xiaotian Han. 2025b. 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. 2025. Data mixing laws: Optimizing data mixtures by predicting language modeling performance.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*.

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.

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.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024b. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *arXiv preprint arXiv:2409.02897*.

Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024c. Longreward: Improving long-context large language models with ai feedback. *arXiv preprint arXiv:2410.21252*.

Junhao Zhang, Richong Zhang, Fanshuang Kong, Ziyang Miao, Yanhan Ye, and Yaowei

Zheng. 2025. 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 Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. 2024d. ∞bench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*.

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

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö. Arik. 2024f. Chain of agents: Large language models collaborating on long-context tasks.

Liang Zhao, Tianwen Wei, Liang Zeng, Cheng Cheng, Liu Yang, Peng Cheng, Lijie Wang, Chenxia Li, Xuejie Wu, Bo Zhu, Yimeng Gan, Rui Hu, Shuicheng Yan, Han Fang, and Yahui Zhou. 2024. Longskywork: A training recipe for efficiently extending context length in large language models.

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).

Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, et al. 2024. Dape: Data-adaptive positional encoding for length extrapolation. *Advances in Neural Information Processing Systems*, 37:26659–26700.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*.

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.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2023. Pose: Efficient context window extension of llms via positional skip-wise training. *arXiv preprint arXiv:2309.10400*.

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.