The Art of CHUNKING in AI Systems:

FOUNDATION AND CORE TECHNIQUES

Master how chunks make intelligence scalable, efficient and powerful.

Dr. Navdeep Singh
Mr. Bavalpreet Singh

The Art of Chunking in AI Systems:

Foundation and Core Techniques

Parab Publications
Sharing Knowledge, Shaping Future

India | UAE | Nigeria | Uzbekistan | Montenegro | Iraq | Egypt | Thailand | Uganda | Philippines | Indonesia

www.parabpublications.com

The Art of Chunking in AI Systems:
Foundation and Core Techniques

Authored By:

Dr. Navdeep Singh
Associate Professor, CSE, Punjabi University,
Patiala, Punjab, India
Email: navdeepsony@gmail.com

Mr. Bavalpreet Singh
AI Architect, CloudCosmos
Toronto, Canada
Email: bavalpreetsinghh@gmail.com

First Impression: August 2026

The Art of Chunking in AI Systems: Foundation and Core Techniques

ISBN: 978-93-48959-06-5

DOI: https://doi.org/10.5281/zenodo.22657742

Rs. 999/- ($80)

DISCLAIMER

Information contained in this book has been published by Parab Publications and has been obtained by the authors from sources believed to be reliable and correct to the best of their knowledge. The authors are solely responsible for the contents of the articles compiled in this book. Responsibility of authenticity of the work or the concepts/views presented by the author through this book shall lie with the author and the publisher has no role or claim or any responsibility in this regard. Errors, if any, are purely unintentional and readers are requested to communicate such error to the author to avoid discrepancies in future.

Published by:
Parab Publications

Preface

The rapid advancement of Large Language Models (LLMs) has transformed the way machines process, understand, and generate human language. However, despite their remarkable capabilities, these models are constrained by context-window limitations and the quality of the information available during inference. Retrieval-Augmented Generation (RAG) has emerged as one of the most influential techniques for overcoming these limitations by combining the reasoning ability of LLMs with external knowledge sources. At the heart of every effective RAG system lies a fundamental yet often underestimated process called **chunking**.

Chunking is far more than simply dividing a document into smaller pieces. It is a critical design decision that directly influences retrieval accuracy, contextual relevance, response quality, computational efficiency, and ultimately the success of an AI application. Whether developing enterprise knowledge assistants, intelligent search engines, document question-answering systems, legal research platforms, healthcare applications, or coding assistants, selecting an appropriate chunking strategy can significantly improve system performance.

Despite its importance, chunking is frequently treated as a minor implementation detail in many tutorials and research discussions. Most existing resources introduce only basic techniques without explaining the underlying principles, trade-offs, or practical considerations required to build robust retrieval systems. This book was written to bridge that gap.

The Art of Chunking in AI Systems: Foundation and Core Techniques provides a comprehensive introduction to the theory and practice of chunking within Retrieval-Augmented Generation systems. Beginning with the fundamentals of RAG, the book explains why retrieval remains essential in the era of increasingly capable language models and demonstrates how chunking serves as the foundation of effective information retrieval.

The chapters explore the taxonomy of chunking methods, covering both fixed-length and dynamic approaches. Readers will learn about character-based chunking, token-based chunking, sliding-window techniques, semantic chunking, document-aware strategies, code-specific chunking, hierarchical chunking, and emerging agentic chunking methods. Rather than presenting these techniques in isolation, the book examines their strengths, limitations, implementation considerations, and appropriate use cases.

A central theme throughout this book is that **there is no universally optimal chunking strategy**. Every application has unique requirements, and the effectiveness of a chunking method depends on the nature of the data, retrieval objectives, embedding models, and downstream language models. Consequently, the final chapters provide practical guidance for selecting, evaluating, and refining chunking strategies for real-world AI systems.

This volume serves as the first part of a broader exploration of chunking technologies. It establishes the conceptual foundation necessary to understand more advanced techniques, evaluation methodologies, optimisation strategies, and future developments that will be presented in subsequent volumes.

The intended audience includes undergraduate and postgraduate students, researchers, educators, software engineers, machine learning practitioners, data scientists, and AI developers who seek a deeper understanding of one of the most important yet overlooked components of modern retrieval systems.

It is our hope that this book enables readers to move beyond treating chunking as a simple preprocessing step and instead appreciate it as a strategic design decision that shapes the intelligence, efficiency, and reliability of AI systems.

Acknowledgment

Writing a book is never an individual effort. It is the culmination of knowledge gained from teachers, researchers, colleagues, students, and the broader scientific community. I would like to express my sincere gratitude to everyone who has directly or indirectly contributed to the completion of this book.

First and foremost, I thank the Almighty for granting me the strength, patience, and perseverance to undertake and complete this work.

I am deeply grateful to the global research community working in Artificial Intelligence, Natural Language Processing, Information Retrieval, and Large Language Models. Their pioneering research, open-source contributions, and scholarly publications have laid the foundation upon which this book has been developed. The collective efforts of researchers around the world continue to advance the field and inspire new ideas.

I extend my heartfelt appreciation to my colleagues and fellow academicians for their encouragement, insightful discussions, and constructive feedback. Their perspectives have enriched my understanding of Retrieval-Augmented Generation and the evolving landscape of AI systems.

My sincere thanks go to my students, whose curiosity, thoughtful questions, and enthusiasm for learning have been a constant source of motivation. Teaching and interacting with them has reinforced the importance of presenting complex concepts in a clear, structured, and practical manner.

I am especially grateful to the developers and contributors of the open-source AI ecosystem. The availability of high-quality frameworks, libraries, benchmark datasets, and educational resources has significantly accelerated innovation and made advanced AI technologies accessible to researchers and practitioners across the world.

I also acknowledge the authors, publishers, and organisations whose books, research papers, technical reports, and documentation have

contributed to my understanding of chunking strategies, retrieval systems, and modern AI architectures. Every effort has been made to present the material in an original, educational, and well-organised manner while respecting the contributions of prior research.

Finally, I owe my deepest gratitude to my family for their unwavering support, understanding, encouragement, and patience throughout the writing of this book. Their confidence in my work provided the motivation to complete this project.

It is my sincere hope that **The Art of Chunking in AI Systems: Foundation and Core Techniques** serves as a valuable resource for students, researchers, educators, and AI practitioners. If this book helps readers better understand the principles of chunking and inspires them to build more effective Retrieval-Augmented Generation systems, then its purpose will have been fulfilled.

Dr. Navdeep Singh and Mr. Bavalpreet Singh

Table of Contents

ABOUT THE AUTHORS

Dr. Navdeep Singh holds a Ph.D. in Computer Engineering from Punjabi University, Patiala, India, and a Master of Engineering degree from Thapar University, Patiala, India. He began his academic career as an Assistant Professor at Punjabi University, Patiala in 2011 and is currently serving as an Associate Professor in the Department of Computer Science and Engineering. With more than 15 years of teaching and research experience, he has been actively involved in undergraduate, postgraduate, and doctoral education, mentoring students and conducting research in emerging areas of computer science. His research interests include Machine Learning, Deep Learning, Natural Language Processing (NLP), Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), and Generative Artificial Intelligence. His current work focuses on developing intelligent AI systems, multilingual NLP, and advanced retrieval techniques for large-scale knowledge systems. Dr. Singh has published more than 30 research papers in reputed national and international journals and conferences. His research contributions reflect a strong commitment to advancing the fields of artificial intelligence and data-driven computing through both theoretical and applied research.

Mr. Bavalpreet Singh is an AI Architect at CloudCosmos, where he designs and deploys production-scale Retrieval-Augmented Generation (RAG) systems, agentic AI platforms, and conversational AI solutions for enterprise clients in the financial and regulatory technology space. With over more than seven years of hands-on experience building intelligent systems, he has led the development of multilingual conversational agents, large language model integrations, and end-to-end AI pipelines used by public-sector and private-sector organizations. He holds a Bachelor of Technology in Computer Science and Engineering from Punjabi University, Patiala, and a postgraduate certification in Artificial Intelligence with Machine Learning (with Honors) from Humber College, Toronto. He is Claude Certified Architect, AWS Certified in Machine Learning Specialty and has published research in conversational AI and named entity recognition. Through his Medium blog and open-source contributions on GitHub, he regularly shares practical engineering insights with the developer community. His work on this book reflects a commitment to bridging the gap between research and real-world AI implementation

ABOUT THE BOOK

The Art of Chunking in AI Systems: Foundation and Core Techniques is a comprehensive guide to one of the most fundamental components of modern Retrieval-Augmented Generation (RAG) systems. While significant attention has been devoted to large language models, embeddings, and vector databases, the process of chunking remains comparatively underexplored despite its critical influence on retrieval quality and AI performance.

This book provides a structured introduction to the principles, techniques, and practical considerations of chunking in AI systems. It begins by establishing the foundations of Retrieval-Augmented Generation, explaining why retrieval continues to play a vital role in enhancing the accuracy, reliability, and contextual understanding of language models.

Readers are then introduced to the complete taxonomy of chunking methods, ranging from traditional fixed-length approaches to advanced dynamic strategies. Topics include character-based chunking, token-based chunking, sliding-window techniques, semantic chunking, document-specific chunking, code-aware chunking, hierarchical chunking, and the emerging concept of agentic chunking. Each technique is presented with its underlying principles, advantages, limitations, suitable applications, and practical recommendations.

Rather than advocating a single "best" method, the book emphasises that effective chunking depends on the characteristics of the data, retrieval objectives, embedding models, and downstream AI applications. It provides readers with a practical framework for selecting, comparing, and evaluating chunking strategies based on real-world requirements.

Designed for students, researchers, educators, AI practitioners, and software developers, this book combines theoretical concepts with practical insights to build a strong foundation in chunking methodologies. Whether you are developing knowledge assistants, enterprise search systems, document question-answering applications, or other RAG-based solutions, this book offers the essential concepts needed to design more accurate, efficient, and scalable retrieval systems.

As the first volume in the The Art of Chunking in AI Systems series, this book lays the conceptual groundwork for advanced topics that will be explored in subsequent volumes, including chunk evaluation, optimisation techniques, adaptive retrieval, multimodal chunking, and next-generation intelligent retrieval architectures.

Parab Publications
Sharing Knowledge, Shaping Future

ISBN 978-93-48959-01-0

India | UAE | Nigeria | Malaysia | Montenegro | Iraq | Egypt | Thailand | Uganda | Philippines | Indonesia

Parab Publications || www.parabpublications.com || info@parabpublications.com