Restored Knowledge Corpus (RKC)

Abstract

As of: Dec 2025

The Restored Knowledge Corpus (RKC) is a large-scale, multi-domain dataset of restored and structured historical texts derived from millions of textbooks. Powered by MonkAI, an automated restoration pipeline, RKC transforms any textbook content into semantically faithful digital editions—preserving text, images, tables, formulas, and layout with machine precision.

Spanning the full breadth of the BISAC classification system, the corpus encompasses thousands of categories across science, technology, engineering, medicine, mathematics, social sciences, and the humanities. Within this breadth, specialized subsets such as STE, Medicine, and Mathematics provide benchmark-quality materials rich in step-by-step reasoning—enabling fine-tuning and evaluation of models designed for long-context comprehension.

Each page achieves semantic and visual fidelity, creating multi-modal data that web-scale corpora cannot match. Beyond digitization, RKC serves as a reasoning-scale infrastructure, connecting centuries of human intellectual effort to the architectures that will define artificial reasoning. In an ecosystem dominated by synthetic and conversational data, it supplies the missing foundation of structured, human-authored knowledge—restored, aligned, and ready for the next generation of AI.