Five-Level Fine-Grained Fully Lossless Knowledge Graph Dataset Verification Report

I. Verification Conclusions

Under current technological constraints, this dataset is verified as: structurally conforming to a standardized “five-level granular knowledge graph,” and content-wise achieving “complete lossless representation of machine-readable information” across textual and tabular dimensions. It is suitable for use as a five-level fully lossless knowledge graph.

II. Structural and Hierarchical Verification

1. Hierarchical Completeness

level 1: chapter

level 2: paragraph

level 3: sentence

level 4: keyword

level 5: media (image/table)

Each level type corresponds one-to-one with its level, with no mixed usage.

2. Correct Parent-Child Relationships

paragraph → Points only to chapter

sentence → Points only to paragraph

keyword → Points only to sentence

media → Points only to chapter

No instances of parent_id pointing to non-existent nodes were found, nor were any circular references detected. The structure tree is well-closed.

III. Content Completeness and “Losslessness”

1. Text Components

Chapters, paragraphs, and sentences are decomposed into independent nodes, each recording page and chapter_title. This enables tracing back to the original text location from any node.

Keywords are automatically extracted from sentence-level content, enhancing retrieval and thematic analysis capabilities without diminishing original information.

→ No substantive loss of textual information at character or contextual levels.

2. Tables

All recognized tables are recorded in media.extra.tables:

caption: Table title;

raw_lines: Raw text of table body line-by-line;

data_points: {row_index, row_text} structure per row.

→ As long as the PDF contains machine-readable table rows, this dataset preserves them line-by-line, enabling complete table reconstruction within the graph.

3. Images and Page Context

For image-based media, the first few lines of text from the containing page are stored in extra.page_excerpt to provide contextual clues;

Pure pixel information in PDFs lacking a text layer (requiring OCR) represents an external technical limitation, not a loss inherent to the graph itself.

IV. Comprehensive Evaluation

Structural Dimension: Five-level hierarchy is clear, parent-child relationships are correct → Five-level granularity is qualified.

Textual Dimension: Chapter → Paragraph → Sentence → Keyword are all traceable → Nearly verbatim lossless.

Tabular Dimension: Table headers + row-level original text are fully stored in raw_lines/data_points → At the textual level, it can be regarded as a lossless mirror image.

Simple Conclusion:

Under the premise of “all machine-readable text within PDFs as the baseline,” this dataset can be regarded as a Level 5 completely lossless knowledge graph. It is suitable for high-level qualitative research, knowledge modeling, and large model integration experiments.