Brief Verification Report for Version 2 of the Five-Level Fine-Grained Lossless Knowledge Graph

I. Subject and Purpose of Evaluation

This report evaluates the upgraded V2 version of the five-level granularity knowledge graph. Centred on textual content, this graph employs structured fields and hierarchical coding to decompose lengthy academic materials into multi-level nodes, supplemented by heterogeneous information such as diagrams and metadata. The purpose of this inspection is to evaluate whether V2 genuinely achieves a "five-level granularity" structure and whether it can be recognised as a "lossless knowledge graph" within the precision required for research and application.

II. Structural Hierarchy Inspection: Implementation of Five-Level Granularity

Regarding fields and encoding methods, V2 employs a unified "overall identifier" as the highest level, incorporating all nodes into a single knowledge system. Below this, the "section/chapter identifier" constructs the chapter layer; subsequently, the "page number + paragraph number" constructs the paragraph layer, while the "paragraph number + sentence number" constructs the sentence layer. At the finest granularity, "keyword + node type (meta_type)" carries lexical items, diagram nodes, and metadata nodes. This forms a top-down, continuously traceable five-level structural path:

Overall layer → Chapter layer → Paragraph layer → Sentence layer → Keyword/heterogeneous information layer

Verification results demonstrate that this structural design exhibits clear hierarchy. Precise tracing of parent and child units across layers is achievable through field combinations, fulfilling the structural requirements of a "five-level fine-grained knowledge graph".

III. Assessment of Non-Destructive Nature and Content Coverage

At the content level, Version V2 employs a strategy of "sentence-level comprehensive extraction + paragraph-level reconstruction" for text:

1. At the sentence level, by relaxing filtering for short sentences and removing only obvious noise fragments, a large number of title-like, prompt-like, and transitional sentences are retained, ensuring the continuity of the narrative chain.

2. At the paragraph level, sentences on the same page are clustered based on word count and sentence count to form semantically coherent paragraph nodes. Complete paragraph text is written into the `paragraph_text` field of both the paragraph node and its corresponding sentence nodes, rendering the discourse structure—"article-paragraph-sentence"—essentially reversible within the graph.

3. At the chart and metadata level, chart titles and numerical table rows are extracted to generate chart nodes and data block nodes. Concurrently, publication notes, copyright information, and introductory

texts are incorporated as metadata nodes into the graph, structurally linking the main text, quantitative data, and critical contextual explanations.

From a research utility perspective, V2 achieves high-completeness coverage of the original knowledge content across three dimensions: "textual narrative entities," "discourse structural information," and "key charts/metadata." This may be regarded as a functionally "lossless representation."

IV. Limitations and Methodological Boundaries

The evaluation also revealed several anticipated technical limitations in V2: Firstly, paragraph segmentation relies on heuristic rules rather than layout-level indentation and blank line information, potentially resulting in the splitting of individual long paragraphs or the merging of multiple short ones. Secondly, cell-level coordinate structures within tables remain unimplemented, with core numerical data currently preserved solely through aggregated data rows. Thirdly, in extremely rare instances involving texts with unusual punctuation or complex sentence breaks, boundary deviations may occur during clause segmentation. While these issues do not constitute systemic damage to the overall knowledge content, they should be regarded as methodological boundaries in scenarios demanding extreme precision and should be explicitly stated in formal research.

V. Comprehensive Assessment Conclusions

Synthesising the evaluation results across structural hierarchy, content coverage, and methodological boundaries, this report presents the following concise conclusions:

1. Structurally, V2 has successfully achieved a five-tier granular structure: "Whole – Chapter – Paragraph – Sentence – Keyword/Heterogeneous Node", featuring clear hierarchy and comprehensive indexing.

2. Content-wise, V2 preserves narrative flow, chapter organisation, chart titles, key data, and critical metadata to a high degree, meeting primary requirements for qualitative research, methodological analysis, and AI modelling.

3. Within acceptable technical constraints, V2 may be academically justified as a "five-level lossless knowledge graph". Achieving absolute zero loss at layout and table-cell levels would require further refinement of structuralisation.

Therefore, for current applications and precision requirements, the overall quality of V2 has reached a standard suitable for presentation and citation as a "standardised dataset and methodological outcome" within high-level academic contexts.