GraphRAG 与多模态知识图谱研究与应用

GraphRAG and Multimodal Knowledge Graphs Research and Applications

中英文逐节对照研究报告
Chinese English Parallel Research Report

研究主线 知识构建 证据检索 可靠生成 多模态应用

核心来源 微软研究院 GraphRAG 文章与美团菜品知识图谱文章

资料检索截止 2026 年 10 月 19 日 / Evidence cutoff October 19 2026

编制完成 2026 年 10 月 19 日 / Completed October 19 2026

正文统计 中文汉字 39,598 ＋ 英文单词 20,577 ＝ 60,175

统计范围含摘要、16 章及两项附录的对照正文，不含目录、标题、图注与书目。英文按词计数，中文仅计汉字。
Count includes paired abstract, chapters and appendices; excludes headings, contents, captions and references.

研究判断应以任务、证据与维护条件为依据。图结构和多模态信息的作用，通过可追踪结果与可检验实验来评估。

Evaluate relationships and modalities through traceable evidence, testable experiments, and sustainable operation.

目录与导航

点击条目跳转到章节；章节末提供返回目录。 / Select an entry to navigate.

研究摘要 Research Abstract

01 研究问题与证据方法 Research Questions and Evidence Method

02 知识图谱的语义与证据基础 Semantic and Evidential Foundations of Knowledge Graphs

03 检索增强生成的机制与限制 Mechanisms and Limitations of Retrieval Augmented Generation

04 微软 GraphRAG 的工作机制 Working Mechanisms of Microsoft GraphRAG

05 图增强检索方法的比较与选择 Comparing and Selecting Graph Augmented Retrieval Methods

06 多模态知识图谱的表示与对齐 Representation and Alignment in Multimodal Knowledge Graphs

07 美团菜品图谱与 CMIngre 案例分析 The Meituan Dish Graph and CMIngre Case Study

08 统一证据架构与工程实现 A Unified Evidence Architecture and Engineering Design

09 分层评测实验设计与不确定性 Layered Evaluation Experimental Design and Uncertainty

10 成本性能与持续运行 Cost Performance and Continuous Operation

11 权限治理安全与责任 Permissions Governance Security and Responsibility

12 企业与科研应用蓝图 Enterprise and Scientific Application Blueprints

13 教育文化与视觉思维实践 Education Culture and Visual Thinking Practice

14 校园菜品多模态 GraphRAG 应用设计 A Campus Dish Multimodal GraphRAG Blueprint

15 可复算的微型图谱实验 A Reproducible Miniature Graph Experiment

16 研究议程采用路线与综合判断 Research Agenda Adoption Roadmap and Synthesis

附录 A 研究与实施操作表 Appendix A Research and Implementation Worksheets

附录 B 双语术语与易混概念 Appendix B Bilingual Terms and Common Confusions

参考文献 References

文献编号 R01 至 R33 在正文中可点击跳转，书目条目包含原始链接。 / Citation identifiers link to the bibliography.

图表索引

F01 两篇核心文章的研究位置与本报告提出的连接层 / Anchor studies and the proposed connection layer

F02 证据对象使陈述能够回到来源 / Evidence objects connect claims to sources

F03 不同检索路径应按任务比较 / Compare retrieval routes by task

F04 层次社区示意与解释边界 / Schematic hierarchy and interpretation boundary

F05 索引摘要与面向问题的汇总不是同一步 / Index reports and query synthesis differ

F06 方法比较矩阵 / Method comparison matrix

F07 多模态资源到细粒度证据的层次 / Granularity from resources to evidence

F08 相似、匹配与确认的状态区分 / Similarity, matching and confirmation

F09 CMIngre 来源数量，单位为图像与食材对 / CMIngre pairs by source

F10 数据集细标签数与检测实验类别数需分开 / Fine dataset labels and coarser detection categories

F11 美团原文图 5.1 多模态菜品图谱构建流程 / Original Meituan Figure 5.1

F12 业务指标与百分点变化，不能混作检测成绩 / Reported business metrics and point changes

F13 从资源到回答共享证据标识 / Shared evidence identifiers from resources to answers

F14 沿依赖链更新与撤回 / Update and withdrawal through dependencies

F15 单一总分不能定位失败来源 / One score cannot locate failures

F16 用对照检查视觉信息是否真正被使用 / Controls for actual visual contribution

F17 用户反馈到组件修复的闭环 / Feedback to component repair

F18 权限应覆盖派生摘要而不只覆盖原文 / Permissions must cover derived reports

F19 同一平台需要不同场景的验收标准 / Shared platform, domain specific acceptance

F20 课堂学习以证据组织与纠错为中心 / Learning through evidence and correction

F21 版本化菜品模型 / Versioned dish model

F22 人工模拟资料中的冲突与未知 / Conflict and unknown in synthetic records

F23 固定成本与边际成本的假设情景 / Assumed fixed and marginal costs

F24 阶段间设置验收与退出条件 / Acceptance and exit conditions between stages

研究摘要

Research Abstract

本报告围绕 GraphRAG 与多模态知识图谱的研究和应用，连接微软研究院关于私有叙事资料发现的 GraphRAG 文章，以及美团关于多模态食材抽取与菜品图谱的技术文章，并结合后续论文、数据集与官方文档进行叙述性综合。报告将知识构建、证据检索和答案生成作为相互依赖但不同的环节，比较社区摘要、关联式记忆、片段组织、文本属性图理解和多模态检索等技术路线。研究重点不是寻找一种普遍最优架构，而是明确关系结构与视觉证据在什么条件下提供可验证价值。

综合分析形成四项判断。第一，图结构能够组织跨片段联系，但效果依赖实体身份、关系类型和任务需求。第二，多模态信息能够补充观察，却不能自动将相似度转化为事实。第三，溯源、版本、权限和不确定性必须贯穿构图到生成，事后附加引用不足以建立可靠性。第四，采用决策需要同时检查任务收益、失败模式和持续维护成本。报告提供统一证据架构、分层评测设计、企业与教学蓝图、校园菜品方案，以及可复算的微型实验。实验仅验证确定性查询与算术，不声称复现大型模型性能。全文按中文与英文逐节对照，图表区分来源数据、原创设计和假设情景。

This report studies GraphRAG and multimodal knowledge graphs by connecting Microsoft's article on discovery in private narrative data with Meituan's article on ingredient extraction and dish graphs, supplemented by later papers, datasets, and official documentation. It treats construction, evidence retrieval, and generation as distinct but interdependent stages and compares community summaries, associative memory, chunk organization, textual graph understanding, and multimodal retrieval. The objective is to identify conditions under which relationships and visual evidence provide verifiable value rather than select one universally optimal architecture.

The synthesis yields four judgments. Graph structure organizes cross passage connections, but value depends on identity, relation types, and tasks. Modalities complement observations without turning similarity into fact. Provenance, versions, permissions, and uncertainty must span construction through generation; adding citations afterward is insufficient. Adoption requires task benefits, failure analysis, and sustainable maintenance together. Outputs include an evidence architecture, layered evaluation, enterprise and teaching blueprints, a campus dish design, and a reproducible miniature experiment. The experiment verifies deterministic queries and arithmetic rather than large model performance. Chinese and English sections are paired, and figures distinguish source data, original designs, and assumed scenarios.

返回目录 Back to contents

第1章　研究问题与证据方法

Research Questions and Evidence Method

研究对象与问题层次

Research Objects and Levels of Questions

本报告研究知识怎样组织，才能让生成式系统在复杂资料中找到依据、连接关系并利用视觉信息。GraphRAG 关注检索与生成过程中图结构的作用，多模态知识图谱关注不同模态怎样指向共同实体、属性和关系。二者相交，却处在不同的分析层次。一个系统可以拥有多模态图谱而不提供生成式问答，也可以使用文本图谱开展 GraphRAG。只有明确输入、处理与输出，才能判断某项改进究竟来自信息增加、关系组织、检索策略，还是生成模型自身的变化。[R07][R16]

研究问题分为事实查找、关系关联和整体理解。事实查找通常寻找某个对象的明确属性，关系关联需要连接多条记录，整体理解则要求概括资料集合的主要议题。它们对证据的要求不同，适合的索引和评价方法也不同。把所有问题放入一个平均分，可能掩盖系统在某类问题上的退步。本报告据此讨论任务与技术的对应关系，而不先假定一种架构能满足全部需求。应用层面的重点，是回答是否适用于当前问题、能否回溯来源，以及维护成本是否与获得的收益相称。

This report examines how knowledge should be organized so that a generative system can locate supporting evidence, connect relationships across complex material, and make use of visual information. GraphRAG concerns the role of graph structure in retrieval and generation. A multimodal knowledge graph concerns how different modalities refer to shared entities, attributes, and relationships. The two subjects intersect, but they belong to different analytical levels. A system can contain a multimodal graph without offering generative question answering. It can also perform GraphRAG over a graph constructed entirely from text. The inputs, transformations, and outputs must therefore be identified before an improvement can be attributed to additional information, relational organization, retrieval strategy, or a change in the generator itself. [R07][R16]

We distinguish factual lookup, relational association, and understanding of a collection as a whole. Factual lookup typically seeks an explicitly recorded property. Relational association requires information from several records to be connected. Collection-level understanding requires a representative account of the main issues across a corpus. These tasks demand different evidence and can justify different indexes and evaluation methods. A single average score can conceal deterioration on one task while another improves. Consequently, the report examines the correspondence between tasks and technical choices rather than assuming that one architecture satisfies every requirement. At the application level, the central questions are whether an answer is applicable to the present request, whether its supporting material can be traced, and whether the cost of maintaining the system is justified by the benefit obtained. These questions determine the unit of analysis throughout the report.

两篇主线文献的不同位置

The Distinct Roles of the Two Anchor Sources

微软 2024 年 2 月的文章提出，用大模型从资料中构建实体关系图，再组织语义群组，以改善分散信息关联和全局主题概括。文章所称私有数据，侧重模型训练时未见过的资料，不等同于所有材料都具有保密属性。其示例以新闻集合说明检索困难，因此应把文章读作一种技术思路及初步证据的介绍。后续论文与官方文档补充了流程，但这些后来的细节不能全部倒填为早期文章已经完成的实验。[R01][R02]

美团文章的主线是食材粒度的中餐理解。它把图像与文字作为互补证据，讨论数据标注、食材检测、跨模态检索和业务知识融合。其学术基准与业务知识图谱指标处于不同层次，不能把某个商品识别错误率当作食材检测准确率。两篇文章共同提醒我们，知识图谱的价值要通过具体信息缺口来判断。本文将前者作为知识使用的案例，将后者作为知识构建的案例；把二者组合成统一系统，则属于本文提出的架构分析，而非两篇原文已经验证的联合成果。[R03][R04]

Microsoft's February 2024 article describes using a language model to construct an entity and relationship graph and organize semantic groups, with the aim of improving association across dispersed information and summarization of corpus-wide themes. Its use of private data emphasizes material that the model did not encounter during training. It does not imply that every example is confidential. The article uses a news collection to illustrate retrieval difficulties and should be read as an introduction to a technical approach accompanied by preliminary evidence. A later research paper and official documentation provide additional implementation detail. Those later details should not be retrospectively presented as experiments already completed in the earlier announcement. [R01][R02]

The Meituan article follows a different line of inquiry: understanding Chinese food at the ingredient level. Images and text provide complementary evidence, and the work discusses annotation, ingredient detection, cross-modal retrieval, and fusion into business knowledge. Its academic benchmark and its business knowledge-graph measurements operate at different levels. An error rate for identifying equivalent products cannot be relabeled as ingredient-detection accuracy. Together, the two sources suggest that the value of a knowledge graph should be assessed against a specific information gap. This report treats the Microsoft work as a case of using organized knowledge and the Meituan work as a case of constructing knowledge from complementary observations. A unified system combining these ideas is an architectural proposal developed in this report, rather than a jointly validated outcome of the two sources. That distinction is retained in the diagrams, examples, and application recommendations.

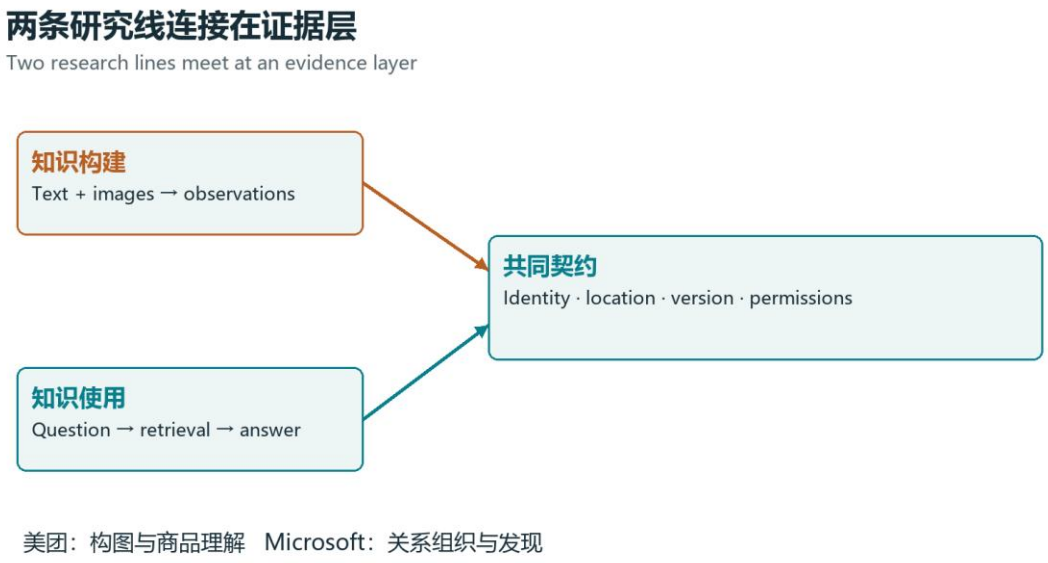


图 01 两篇核心文章的研究位置与本报告提出的连接层 / Anchor studies and the proposed connection layer
来源 / Source 本报告原创设计 / Original design

检索路径与纳入原则

Search Routes and Inclusion Principles

本次检索采用结构化叙述综述的方法，围绕 GraphRAG、检索增强生成评价、多模态知识图谱、跨模态实体对齐和食材理解形成主题组。Consensus 用于发现图检索与评价研究，并在引用检索结果前读取论文记录；SciSpace 用于语义检索相关多模态研究及补充方法信息。对于决定技术判断的条目，再查作者论文、正式出版页面、项目文档或数据集说明。检索入口提供的题名、年份与摘要不是最终事实来源，尤其要检查预印本年份、期刊卷期和版本修订是否混在一起。[R07][R09][R26]

纳入原则是与问题直接相关、来源可追溯、方法或评价对象能被辨认。排除仅重复宣传语、缺乏原始依据或无法确认对象的内容。对于只读到摘要或出版记录的研究，报告只采用相应层次能够支持的概括，不推断未阅读的消融实验或数据处理细节。检索截止日为 2026 年 10 月 19 日。由于没有开展穷尽数据库检索、双人独立筛选和完整偏倚评价，本文不自称系统综述，也不报告虚构的筛选数量或统计合并结果。文献清单和读取范围随报告保留，便于后续复查与更新。

The review uses a structured narrative-synthesis approach. Searches were organized around GraphRAG, evaluation of retrieval-augmented generation, multimodal knowledge graphs, cross-modal entity alignment, and ingredient-level food understanding. Consensus was used to discover graph retrieval and evaluation studies, and the corresponding paper records were fetched before any search result was cited. SciSpace supported semantic searches for multimodal work and supplementary methodological information. Items that affected technical judgments were then checked against author papers, publisher records, official project documentation, or dataset descriptions. Titles, years, and abstracts returned by a search interface are discovery aids rather than the final authority. Particular care is needed when an interface mixes the year of a preprint, a later journal volume, and subsequent revisions. [R07][R09][R26]

Studies were included when they directly addressed a research question, had a traceable source, and described an identifiable method or evaluation object. Material consisting only of promotional repetition, unsupported claims, or an unclear research object was not used as substantive evidence. When access was limited to an abstract or a publication record, the report uses only the high-level statements supported by that material. It does not infer unexamined ablations or preprocessing details. The search cutoff is 19 October 2026. This work did not conduct an exhaustive search of every database, independent screening by two reviewers, or a complete risk-of-bias assessment. It therefore does not claim to be a systematic review, and it does not invent screening counts or pooled statistical effects. The bibliography and recorded reading scope are retained so that the synthesis can be checked and extended.

证据强度与版本识别

Evidence Strength and Version Identification

本报告采用四类证据标签。原始研究结论指作者在指定数据、模型和评价条件下报告的发现，保留实验范围。官方实现说明指某个软件或数据集的公开功能与数据结构，保留访问日期。本文分析指对不同来源的比较、解释和设计推导，其合理性需要读者根据论证判断。可复算模拟指用明确输入和规则得到的演示结果，它证明计算过程符合设定，不能证明真实模型在生产环境具有同样表现。这些标签没有排成简单的高低等级，因为它们回答的问题不同。

同一研究的博客、预印本、修订稿与正式出版版本可能存在作者、标题细节或评价设置的变化。报告在需要时同时记录首次公开时间和采用版本，避免把 2025 年的修订结果写成 2024 年 2 月的原始结论。对于官方文档，当前可见功能也不能自动证明论文实验使用了该功能。图示中的关系还需区分来源支持、设计建议、教学模拟与待核查。一个流程按顺序连接，只说明处理先后；它并不自动构成现实因果关系。一个实体与另一实体同处社区，也不意味着两者之间存在直接业务关系。

The report distinguishes four categories of evidence. A reported research finding is a result stated by the authors under specified data, model, and evaluation conditions; its experimental scope is retained. An official implementation description concerns a publicly documented function or data structure and is associated with an access date. Analysis in this report consists of comparison, interpretation, and architectural reasoning developed from the sources; readers must assess the argument rather than treating it as an observed result. A reproducible simulation is an outcome calculated from explicit inputs and rules. It establishes that a calculation follows its assumptions, but does not establish that a real model will behave the same way in production. These categories are not arranged as a simple hierarchy, because they answer different kinds of questions.

A research announcement, initial preprint, revised manuscript, and formal publication can differ in authorship, title details, or experimental conditions. Where relevant, the report records both the first public date and the version being used. This prevents results added in a 2025 revision from being attributed to a February 2024 announcement. Similarly, a capability currently present in official software documentation does not establish that the capability was used in the paper's experiments. Relationships in figures also distinguish source-supported information, proposed design, simulation, and unresolved evidence. A sequential connection in a process diagram describes an order of processing. It does not by itself establish real-world causation. Two entities located in the same graph community do not necessarily

have a direct business relationship. Such distinctions are essential when a visually persuasive diagram could otherwise make a qualified proposition appear settled.

中英对照与视觉表征规则

Rules for Bilingual Alignment and Visual Representation

中英对照按同一节的论点与证据配对，保持数字、单位、否定条件和不确定性一致。本文将 grounding 译为依据约束或证据落地，强调输出与可定位材料之间的联系；将 faithfulness 译为忠实性，强调回答是否准确反映给定材料；将 factual correctness 译为事实正确性，强调命题与现实或可靠事实标准是否一致。三者有关，但不互相替代。错误材料可以被忠实转述，有引用的答案也可能引用不当。翻译时如果把“有依据”统一改成“真实可靠”，就会扩大原文的承诺。

图表根据要解释的问题选择。流程图表达处理顺序与信息传递，带标签的关系图表达实体连接，矩阵用于保持比较维度一致，数量图用于呈现真实测量或标明假设的计算。图中的颜色辅助辨认对象类型，不能独立承担结论。所有数量图保留输入和计算口径，不给不同任务的百分比拼造一个综合得分。大模型生成的说明文字属于待核查的中间产物，不能因为排版成图就获得事实地位。图源与数据随报告提供，使读者能够修改表达并检查关系是否仍成立。

Chinese and English passages are paired around the same argument and evidence. Numbers, units, negative conditions, and degrees of uncertainty are preserved across the two versions. Grounding refers here to constraining an output by material that can be located and examined. Faithfulness refers to whether the answer accurately represents the supplied material. Factual correctness concerns whether a proposition agrees with reality or an appropriate factual standard. These properties are related, but they are not interchangeable. Incorrect source material can be reproduced faithfully, and an answer containing citations can still use those citations improperly. Translating all of these concepts as a broad guarantee of truth would strengthen the claims beyond their evidence.

Visual forms are selected according to the question being explained. Process diagrams represent sequence and the transfer of information. Labeled relationship graphs represent connections among entities. Matrices preserve common comparison dimensions. Quantitative charts display observations or calculations whose assumptions are stated. Color helps distinguish object types and does not carry the conclusion by itself. Every quantitative figure retains its inputs and measurement convention. Percentages from different tasks are not combined into an invented overall score. Explanations generated by a language model remain intermediate material requiring examination; placing them inside a diagram does not turn them into established facts. The figure sources and associated data are supplied with the report so that readers can revise the presentation and examine whether the underlying relationship still holds. These rules make the illustrations part of the argument rather than decoration added after the analysis.

报告的决策用途与研究边界

Decision Uses and Boundaries of the Report

本报告面向需要理解、研究或设计知识增强系统的读者，提供三个层面的支持。概念层帮助区分图谱、检索、跨模态对齐和生成，避免仅根据工具名称判断能力。设计层说明数据来源、索引、

证据组织、模型调用与维护之间怎样配合。评价层提出可观察的质量标准和实验控制，使方案可以被证伪，而不只被演示。这里的应用建议属于研究与工程设计建议，不能代替对具体机构数据、人员、业务约束和可接受风险的实地调查。

报告不声称训练了新的模型，也不声称复现了所有文献的性能。执行部分限于透明的微型图查询、指标复算和合成成本分析，目的是让读者看见概念差异与计算约束。企业、教育、文化与餐饮案例用于展示迁移方法，除注明引用的原文业务结果外，其系统方案均为本文设计。读者可以先建立一个小而完整的证据流程，确认它能处理正常、缺失和冲突资料，再决定是否增加复杂模型。研究价值体现在假设能否被检验、失败能否被定位、结论能否被其他人检查，而不体现在图的规模或术语的数量。

The report is intended for readers who need to understand, investigate, or design systems that augment generation with organized knowledge. It supports decisions at three levels. At the conceptual level, it separates graphs, retrieval, cross-modal alignment, and generation so that capability is not inferred from a product name alone. At the design level, it explains how sources, indexes, evidence organization, model calls, and maintenance interact. At the evaluation level, it proposes observable quality criteria and experimental controls that allow a design to be challenged rather than merely demonstrated. Application recommendations are research and engineering proposals. They do not replace investigation of a particular institution's data, staff, operational constraints, and acceptable risks.

The report does not claim to train a new model or reproduce the performance of every cited study. Executed work is limited to transparent miniature graph queries, verification of numerical metrics, and synthetic cost calculations. These activities make conceptual distinctions and computational constraints visible. Enterprise, educational, cultural, and food-service examples illustrate transfer to new settings. Except for explicitly attributed business measurements, the proposed systems in those examples are designs developed for this report. Readers can begin with a small but complete evidence workflow, establish that it handles ordinary, missing, and contradictory information, and then decide whether additional model complexity is warranted. Research value depends on whether assumptions can be tested, failures can be localized, and conclusions can be inspected by others. Graph size and the number of technical terms provide no substitute for those properties.

阅读路径

Reading Routes

概念学习可按第一至第七章阅读，先理解资料与知识的区别，再比较方法并进入菜品案例。准备项目的读者可重点阅读第八至第十二章以及第十四章，结合架构、评测、成本和治理形成试点方案。教学使用可从第十三章与第十五章开始，以证据卡和微型实验引出理论。第十六章提出可检验研究议程，附录提供研究操作表与术语辨析。不同路径共享一个核心要求：每个重要结论都应能回到证据，每个应用承诺都应能转化为可检查任务。

Conceptual readers can follow Chapters 1 to 7, moving from sources and knowledge to methods and the dish case. Project teams can focus on Chapters 8 to 12 and 14 to combine architecture, evaluation, cost, and governance. Teaching can begin with Chapters 13 and 15, using evidence cards and the miniature experiment to introduce theory. Chapter 16 proposes testable research directions, and appendices provide

operational worksheets and terminology. Every route shares one requirement: important conclusions should trace to evidence, and application promises should become inspectable tasks.

返回目录 Back to contents

第2章　知识图谱的语义与证据基础

Semantic and Evidential Foundations of Knowledge Graphs

图结构与知识命题

Graph Structure and Knowledge Claims

图由节点和边构成，但知识图谱还需要说明节点指代什么、边表达什么，以及记录适用于怎样的语境。窗口、菜品和食材可以分别成为实体，供应和包含可以成为有方向的关系。名称只是实体的一种属性，不能承担全部身份。两个同名菜品可能属于不同商户、不同配方或不同时间版本。如果直接把名称相同当作实体相同，图中看似合理的路径就可能连接了并不相同的对象。稳定标识、类型和来源约束，是图谱能够支持可靠检索的前提。[R15][R16]

知识命题与存储形式也应区分。某条记录声称一道菜包含花生，这是一个带来源的陈述；把陈述写成一条边，只改变了表示方式，没有完成事实核验。图谱的优势在于让连接可以被查询、组合和检查，而不是自动把文本转化为真实知识。本报告把实体关系图理解为可计算的主张结构，并要求对关键边记录依据。生成系统从图谱取出路径后，仍须判断路径的关系含义是否符合问题。仅仅存在一条连通路径，不能证明它就是答案所需的推理链。

A graph consists of nodes and edges, but a knowledge graph must additionally explain what its nodes denote, what its edges mean, and in which context a record applies. A service counter, a dish, and an ingredient can be represented as different entities, while serves and contains can be directed relationships. A name is an attribute of an entity rather than a complete identity. Two dishes with the same name can belong to different merchants, recipes, or time periods. If identical names are treated as identical entities, an apparently plausible path can connect objects that should remain separate. Stable identifiers, types, and source constraints are prerequisites for useful graph retrieval. [R15][R16]

The claim represented by a record should also be distinguished from the form in which it is stored. A record stating that a dish contains peanuts is a source-attributed assertion. Encoding it as an edge changes its representation without verifying its truth. A graph makes connections queryable, composable, and inspectable; it does not automatically transform text into established knowledge. We therefore treat an entity relationship graph as a computational structure of claims and require supporting material for important edges. After retrieving a path, a generative system still has to determine whether the meanings of its relationships answer the question. Connectivity alone is insufficient. A path can be mathematically valid while being semantically irrelevant, temporally inconsistent, or supported by evidence that does not apply to the object under discussion.

本体与关系粒度

Ontology and Relationship Granularity

本体为实体类型、关系类型和适用约束提供共享约定。设计时应先从要回答的问题出发，再决定哪些对象必须独立成节点。若只需按菜系筛选，菜系可以作为属性；若需要追踪菜系的分类层级、别名和研究来源，则把它建成实体更便于维护。一个设计没有脱离任务的绝对最优形式。过度细分会增加标注和维护成本，过度概括又可能丢失问题所需的差异。合适的粒度取决于查询、证据和更新责任之间的配合。

多方关系尤其容易被简单二元边误写。商户在某日供应某道特定配方的菜，包含商户、日期、菜品版本和供应状态。若只保存“商户供应菜品”，有效期就会消失。可将一次供应记录或一个菜单版本建为中间对象，使时间和来源依附于具体事件。类似地，图片区域属于某张图片，又被某个模型识别为某种食材，这应与菜品真实配方中的食材关系区分。本文建议在关系模式中明确观察、声称、推测和确认的差别，使视觉识别与业务事实之间保留可检查的转换环节。

An ontology provides shared conventions for entity types, relationship types, and applicable constraints. Its design should begin with the questions to be answered and then determine which objects need independent identity. Cuisine can be stored as an attribute when the only requirement is filtering dishes. If the system must maintain cuisine hierarchies, aliases, and research sources, representing cuisines as entities can be more useful. There is no task-independent optimal representation. Excessive detail increases annotation and maintenance costs, while excessive aggregation removes distinctions needed by the questions. Appropriate granularity depends on the relationship among queries, evidence, and responsibility for updates.

Relationships involving several participants are particularly vulnerable to distortion when compressed into a simple binary edge. A merchant serving a specific recipe on a particular date involves a merchant, a date, a dish version, and a supply status. Storing only merchant serves dish loses the period of validity. A supply record or menu version can instead be represented as an intermediate object to which time and provenance are attached. Likewise, a region belongs to an image and is classified as an ingredient by a particular model. That observation should remain distinct from an ingredient documented in the dish's actual recipe. This report recommends making observation, assertion, inference, and confirmation explicit in the relationship scheme. Doing so preserves an inspectable transition between visual recognition and business facts, rather than silently converting every model prediction into an authoritative property of the world.

来源记录与可追溯链

Provenance Records and Traceable Chains

来源记录应回答一条知识从哪里来、经过什么处理、由谁或什么系统产生，以及何时有效。W3C 的 PROV 框架提供实体、活动和责任主体等概念，帮助表达数据的产生与派生关系。应用不必完整照搬一个标准的全部细节，但需要保留足够的信息，使结论可以回到具体文档、页面、图片区域或原始记录。只保存网站首页链接，通常不足以支持细粒度核查。[R28]

在 GraphRAG 中，原文、抽取关系、社区报告和最终回答构成多层派生链。一个回答引用社区报告时，读者还应能够继续追到报告依据的片段。多模态系统还要记录图片版本、裁切坐标和识别结果，避免更新图片后旧区域失去含义。本文建议为每次处理保存输入标识、模型或规则版本、输出标识和状态。这样，某份材料撤回或纠正时，可以定位受影响的关系、摘要和缓存。来源链提供了审查条件，却不自动保证原始资料可靠；独立事实核验与派生过程审计仍是两个不同任务。

Provenance should explain where a piece of knowledge came from, how it was processed, which person or system produced it, and when it applies. The W3C PROV framework supplies concepts including entities, activities, and agents for describing production and derivation. An application need not adopt every element of a standard, but it should retain enough information for a conclusion to be traced to a document, page, image region, or original record. A link to a website's home page is usually inadequate for fine-grained verification. [R28]

In GraphRAG, source text, extracted relationships, community reports, and final answers form a chain of derived objects. When an answer cites a community report, the reader should be able to continue to the passages supporting that report. A multimodal system must additionally preserve image versions, crop coordinates, and recognition outputs. Otherwise an updated image can leave an old region reference without a meaningful target. We propose recording the input identifier, model or rule version, output identifier, and status for each processing activity. A withdrawn or corrected source can then be connected to the relationships, summaries, and caches that depend on it. Such a chain creates conditions for inspection; it does not establish that the original source is reliable. Independent factual verification and auditing the derivation process remain separate activities. A well-recorded transformation can faithfully preserve an unreliable claim, which is why both activities need an explicit place in the system.

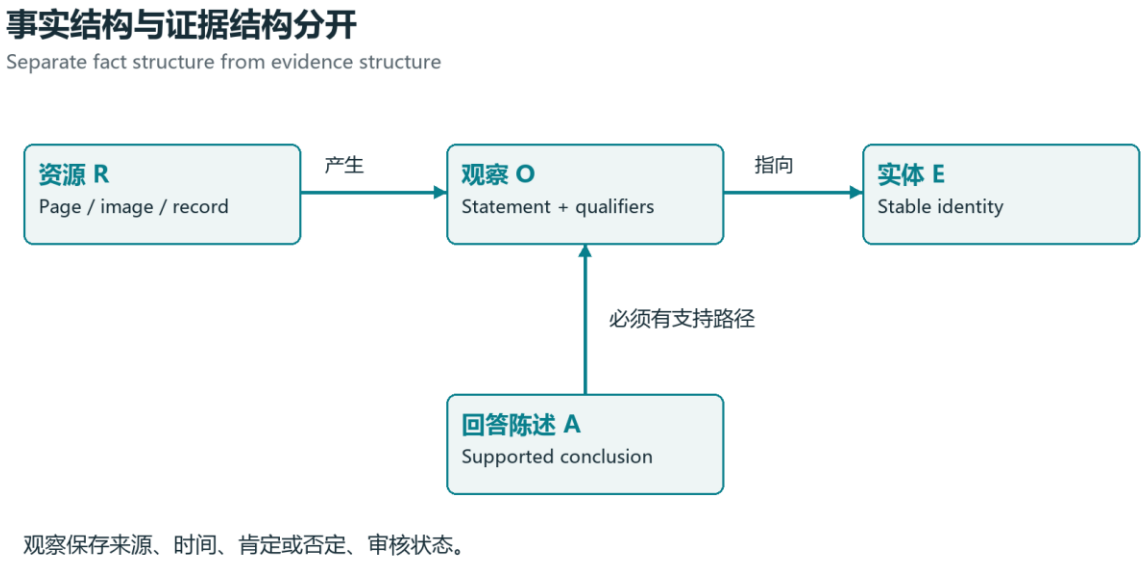


图 02 证据对象使陈述能够回到来源 / Evidence objects connect claims to sources
来源 / Source 本报告原创设计 / Original design

实体身份与歧义处理

Entity Identity and Ambiguity Resolution

实体对齐首先是身份判断，然后才是相似度计算。相同名称可能指向不同对象，不同名称也可能指向同一对象。食材别名、机构简称、人物同名与商品版本分别具有不同的歧义结构，不能共享一条未经验证的合并规则。候选生成可以使用名称、语境、关系邻域或图像特征，但合并决策还应满足业务约束。例如两条记录的有效时间完全冲突，或者一种食材在图中只被概括为“蔬菜”，都不宜因为向量相近就确认细粒度身份。[R16][R22]

本文建议把确定合并、候选关联和明确不同三种状态分开。候选关联可以帮助召回，但不应直接进入需要精确身份的结论。每次合并应保留原始标识与可撤销映射，而不是删除旧记录。拆分和纠错同样重要，因为错误合并会让关系沿着错误身份扩散。跨模态信息可以提供额外线索，也可能引入新歧义：一张图片可能展示多个对象，文字可能只提到其中一个，配图还可能仅作示意。因此，图文配对不等于图中所有对象都与文本实体一一对应。

Entity alignment begins with an identity decision rather than with a similarity calculation. The same name can refer to different objects, and different names can refer to the same object. Ingredient aliases, abbreviated organization names, people sharing a name, and product versions create different ambiguity patterns. They should not be governed by one unvalidated merge rule. Candidate generation can use names, context, relationship neighborhoods, or visual features, but a merge decision should also satisfy domain constraints. Records with incompatible periods of validity should not be merged merely because their embeddings are close. Nor should an image labeled only as vegetables automatically be assigned a specific ingredient identity. [R16][R22]

We recommend separating confirmed identity, candidate association, and confirmed difference. Candidate associations can support recall, but should not silently support conclusions that require exact identity. A merge should preserve original identifiers and a reversible mapping rather than deleting the former records. Splitting and correction deserve equal attention, because an incorrect merge can propagate relationships through the wrong identity. Multimodal information offers additional clues and additional opportunities for ambiguity. An image may depict several objects while the accompanying text mentions only one. An illustration may be representative rather than a photograph of the exact item described. Consequently, an image-text pairing does not imply a one-to-one correspondence between every visual object and a textual entity. An alignment record should state the scope of the correspondence and retain unresolved cases for examination.

存储结构与检索接口

Storage Structures and Retrieval Interfaces

图谱存储、向量索引与原始素材库可以承担互补职责。图结构适合保存明确关系及约束，向量索引适合按语义或视觉相似性生成候选，素材库保留完整文档和高分辨率图片。将这些职责分开并不要求部署三个独立产品，而是要求系统能清楚回答每类数据由谁管理、怎样更新、如何查询。关系数据库也可以表示部分图结构，图数据库也可以提供向量检索。研究比较应关注数据模型和查询行为，不宜仅按照数据库类别推断效果。

检索接口最好返回带类型的证据对象，而不只返回一段拼接文字。证据对象可以包含实体标识、关系、来源位置、时间、权限与不确定状态。生成层据此选择怎样表达结论，审计层据此判断一次回答是否使用了有效材料。对于需要复核的图片，接口还应给出区域与原图的连接，而不是只给一个无法解释的向量分数。本文把证据接口看作架构中的稳定契约：模型和存储可以替换，但证据必须以可核查的形式传递。接口若过早把结构压成自然语言，后续组件将难以恢复丢失的限制条件。

Graph storage, a vector index, and a repository of original assets can perform complementary functions. The graph preserves explicit relationships and constraints. The vector index generates candidates through semantic or visual similarity. The asset repository retains full documents and high-resolution images. Separating these responsibilities does not necessarily require three different products. It requires a clear account of ownership, update behavior, and query semantics for each category of data. A relational database can represent some graph structures, and a graph database can provide vector retrieval. Research comparisons should therefore examine models and query behavior instead of inferring performance from a database category alone.

A retrieval interface should ideally return typed evidence objects rather than only a concatenated block of prose. An evidence object can carry entity identifiers, relationships, source locations, time, permissions, and an uncertainty status. The generation layer can use this information to formulate a qualified conclusion, while the audit layer can determine whether the answer used applicable material. For an image requiring review, the interface should preserve the connection between a region and the original image instead of exposing only an opaque vector score. We treat this evidence interface as a stable architectural contract. Models and storage systems can change, but supporting information must continue to move in an inspectable form. If an interface converts structure into prose too early, downstream components may be unable to recover the constraints that were discarded. Keeping these constraints explicit makes both integration and debugging more reliable.

图谱质量与查询边界

Graph Quality and Query Boundaries

图谱质量至少涉及覆盖、正确性、一致性、时效性和来源完整性。这些维度可能相互牵制。提高自动抽取的召回率可能引入更多错误边，严格过滤又可能遗漏罕见但重要的关系。一个高度连通的图也可能主要由泛化过度的标签组成，难以回答具体问题。因此，节点数、边数和连通率可用于监测数据变化，却不能单独作为知识质量的证明。评价需要从实际查询出发，观察关键对象、关系和证据是否以合适粒度得到保留。

缺失边的解释尤其关键。在开放资料中，未找到某条关系通常表示系统尚无足够记录，并不表示该关系在现实中为假。只有当数据范围、采集完整性与业务规则允许封闭解释时，才能把缺失视作否定。路径长度也不是推理质量的替代指标，较短路径可能语义错误，较长路径可能包含必要的时间或来源条件。本文建议在查询层显式区分肯定、否定、未知和冲突，并让回答层保留这些状态。这样，系统可以诚实说明资料不足，而不把图的空白填充为自信的判断。

Graph quality involves coverage, correctness, consistency, timeliness, and completeness of provenance. These dimensions can pull in different directions. Increasing the recall of automatic extraction may

introduce additional incorrect edges, whereas strict filtering may remove rare but important relationships. A highly connected graph can also be dominated by overly broad labels that are unsuitable for specific questions. Node counts, edge counts, and connectivity can monitor changes in the data, but they cannot independently establish knowledge quality. Evaluation should begin with actual queries and examine whether the relevant objects, relationships, and evidence have been preserved at an appropriate granularity.

The interpretation of a missing edge is especially important. In an open collection of material, failing to find a relationship normally means that the system lacks sufficient records. It does not establish that the relationship is false in the world. Absence can be interpreted as negation only when the data scope, collection completeness, and domain rules justify a closed interpretation. Path length is likewise no substitute for reasoning quality. A short path can be semantically wrong, while a longer one may retain essential time or provenance conditions. We recommend that the query layer explicitly distinguish positive evidence, negative evidence, unknown information, and conflict, and that the answer layer preserve those distinctions. This enables a system to explain evidential insufficiency rather than filling gaps in its graph with confident conclusions. The ability to represent uncertainty is part of the knowledge model, not merely a wording preference at the end of generation.

返回目录 Back to contents

第3章 检索增强生成的机制与限制

Mechanisms and Limitations of Retrieval Augmented Generation

参数知识与外部证据

Parametric Knowledge and External Evidence

语言模型的参数承载训练中形成的统计规律，外部资料则提供可更新、可定位的任务信息。RAG 把检索得到的材料加入生成上下文，使模型能够围绕这些材料回答问题。经典研究将参数化生成与非参数记忆结合，形成知识密集任务的一种方法；这一思想不意味着外部知识被永久写进模型权重，也不意味着模型只能使用给定材料。[R06]

在应用中，需要区分模型可能知道什么与系统允许依据什么回答。面向机构资料的问答，往往要求回答以指定版本的内部记录为准，即使模型能凭一般知识生成一个看似合理的答案。如果资料不足，理想行为可能是说明不足，而不是调用常识填满空白。本文建议将回答依据写成可测试规则，例如关于本机构流程的结论必须附有效文档位置。模型是否遵守该规则，应通过缺失资料、过期文件和冲突记录测试，而不能只用正常问题演示。由此，RAG 的核心控制对象是证据选择及其使用方式。

A language model's parameters encode statistical regularities acquired during training, whereas external material supplies task information that can be updated and located. RAG adds retrieved material to the generation context so that the model can answer with reference to it. The foundational work combines parametric generation with non-parametric memory for knowledge-intensive tasks. This does not imply

that retrieved knowledge is permanently written into the model's weights, nor does it guarantee that the generator will use only the supplied evidence. [R06]

An application must distinguish what a model might know from what the system is permitted to use as the basis of an answer. Question answering over institutional documents often requires conclusions to follow a specified version of internal records, even when general knowledge would allow a plausible response. When the records are insufficient, the desired behavior may be to explain that insufficiency rather than fill the gap with common knowledge. We recommend expressing evidential requirements as testable rules. For example, a statement about an institution's procedure should cite a location in a currently applicable document. Compliance should be examined using missing material, obsolete documents, and contradictory records, not only demonstrated with ordinary questions. The central control problem in RAG is therefore the selection of evidence and the manner in which it is used. This framing also explains why a stronger generator alone cannot repair every retrieval failure.

检索单位与切分策略

Retrieval Units and Segmentation Strategies

切分决定检索器能够返回什么对象。固定长度片段容易实现，但可能把条件与结论、表头与数值、图片与说明分开。按标题、段落或版面区域切分，能够保留更多结构，却仍需处理跨页表格和跨段指代。片段越小，定位通常更精细，但独立理解所需的上下文可能不足；片段越大，信息更完整，却会占用更多上下文并引入无关内容。本文把切分视为证据边界设计，而不是一个只需调节长度的预处理步骤。

可检索对象应保留父文档、相邻片段和结构位置。检索命中一个数值时，系统还要知道对应指标、单位、时间和人群；命中一个图片区域时，也要知道它来自哪一页和哪一版本。图结构能够连接这些对象，但首先必须在数据处理时保留连接所需的标识。对于条款、流程和实验材料，建议建立小片段定位与大范围补充的双层机制，先召回具体证据，再补充解释它所需的局部结构。是否有效应通过答案支持率与上下文成本共同判断。

Segmentation determines which objects a retriever can return. Fixed-length chunks are easy to implement, but they can separate a condition from its conclusion, a table heading from its values, or an image from its explanation. Segmentation by headings, paragraphs, or layout regions retains more structure, but still has to handle tables continuing across pages and references spanning paragraphs. Smaller chunks can provide more precise localization while lacking the context needed for independent interpretation. Larger chunks preserve more information but consume additional context and can introduce irrelevant material. We treat segmentation as the design of evidence boundaries rather than as a preprocessing step controlled only by length.

Retrievable objects should retain their parent document, neighboring passages, and structural location. When retrieval returns a number, the system also needs its metric, unit, period, and population. When it returns an image region, it needs the source page and version. Graph structure can connect these objects only if the identifiers required for those connections were preserved during ingestion. For procedures, clauses, and experimental material, a useful design is to combine fine-grained localization with retrieval of the surrounding structure. The system first identifies specific evidence and then adds the local context required to interpret it. The effectiveness of this design should be evaluated jointly through the support it

provides for answers and the context resources it consumes. A segmentation method that improves one retrieval metric while repeatedly removing essential conditions may be unsuitable for the actual task.

候选召回与排序组合

Candidate Retrieval and Ranking Combinations

不同检索信号可以互补。关键词匹配擅长保留专名、编号和明确表达，向量相似性可帮助找到措辞不同但语义相关的材料，图邻域能扩展到关联对象，时间与权限则限定候选是否可用。组合检索需要说明每个信号承担的职责。把多个分数直接相加并不天然合理，因为量纲、分布和校准方式可能不同。融合方法应以开发集检验，避免根据少数演示问题手工调出一个看似优越的结果。

重排序将较少的候选交给更精细的相关性模型或规则。它可以改善候选顺序，却不能找回召回阶段从未返回的必要证据。若答案依赖两条彼此分散的记录，仅让每个片段分别与问题打分，也可能遗漏它们组合后的价值。本文建议同时记录候选来源、被过滤原因和排序变化，并评价集合是否覆盖完整证据链。与其只追求首条结果相似，不如明确哪些问题需要单条证据、哪些需要关系闭合、哪些需要覆盖多个主题。检索策略由此才能与研究问题保持一致。[R09][R13]

Different retrieval signals can complement one another. Lexical matching preserves names, identifiers, and explicit expressions. Vector similarity helps locate material phrased differently but related in meaning. Graph neighborhoods can expand retrieval to associated objects. Time and permissions constrain whether a candidate is usable. A hybrid design should specify the responsibility of each signal. Simply adding scores does not make them commensurable, because their scales, distributions, and calibration can differ. Fusion should be examined on a development set instead of manually tuned to make a small demonstration look favorable.

Reranking passes a smaller candidate set through a more detailed relevance model or rule. It can improve ordering, but it cannot restore necessary evidence that was never retrieved. If an answer depends on two dispersed records, scoring each passage against the question independently can also miss the value of their combination. We recommend recording where candidates came from, why they were filtered, and how their order changed. Evaluation should examine whether the selected set covers the complete evidence chain. It is insufficient to optimize only the similarity of the first result. Some questions require one explicit record, others require a relationship to be completed, and still others require coverage across several themes. Distinguishing these requirements allows retrieval design to remain aligned with the research question. Comparative GraphRAG studies and graph-guided expansion methods provide reasons to examine this alignment, without establishing one universally optimal fusion rule. [R09][R13]

上下文组织与生成约束

Context Organization and Generation Constraints

检索之后仍有一个独立问题：怎样组织材料，才能让模型正确使用它们。证据顺序、重复内容、来源冲突和长距离依赖都会影响上下文。研究曾观察到，在所测试模型和任务中，相关信息的位置会改变表现，因此“能放入上下文”与“能有效利用”需要分别验证。这个发现不应被扩展为对所有后续模型的固定性能判断，但它提醒我们保留位置与长度的控制实验。[R29]

本文建议把上下文包装成有来源和状态的证据集合，清楚区分原文摘录、机器摘要与推测说明。对需要精确引用的问题，应优先保留支持结论的直接片段；对需要概括的问题，可以提供摘要，但同时保留回溯入口。生成指令应要求在证据不足、条件不匹配或记录冲突时表达相应限制。指令的存在不等于约束已生效，评价必须检查实际输出。一个常见错误是把模型自己的上一轮答案再次作为证据加入上下文，形成自我支持的闭环。若允许使用先前生成内容，就必须保留其派生身份及原始依据。

Retrieval leaves a separate problem unresolved: how should material be organized so that the model uses it appropriately? Evidence order, duplication, conflicting sources, and long-distance dependencies can all affect the context presented to the generator. Research has observed that the position of relevant information changes performance for the models and tasks examined. The ability to fit material into a context window should therefore be evaluated separately from the ability to use it effectively. This observation is not a fixed performance claim about every later model, but it motivates controlled experiments involving position and length. [R29]

We recommend packaging context as a set of evidence objects with explicit sources and statuses. Original excerpts, machine-generated summaries, and speculative explanations should remain distinguishable. Questions requiring precise attribution should retain the passages directly supporting the conclusion. Questions requiring synthesis can use summaries while preserving routes back to their sources. Generation instructions should require appropriate qualification when evidence is missing, conditions do not match, or records conflict. The existence of such an instruction does not establish that it is effective; evaluation must inspect the actual answers. A particularly problematic practice is to reintroduce the model's previous answer as if it were independent evidence, creating a self-supporting loop. If earlier generated material is reused, its derived status and original basis must be retained. Otherwise repetition can make a claim appear corroborated when every instance ultimately comes from the same unsupported generation.

多跳问题与问题路由

Multi Hop Questions and Query Routing

多跳问题要求把若干证据步骤连接起来，但步骤数量与推理难度并不完全一致。一个两跳问题可能需要困难的实体消歧，一个较长路径则可能只是明确的层级展开。系统应辨认所需关系、对象类型和限制条件，而不是机械地增加遍历深度。对于“哪些窗口供应含鸡肉菜品”，需要连接窗口、菜品与食材；若问题再限定日期，还必须加入菜单版本。缺失其中一个条件时，路径即使连通也不完整。

问题路由可以把请求分配给具体查找、关系扩展、全局摘要或跨模态核查等处理方式。路由器本身也会出错，因此要保留回退路径和运行记录。一个稳妥的研究设计应比较固定路由、规则路由和模型路由，并让它们共享相同语料与生成预算。若路由使用了测试集答案或人工事后选择，得到的优势就不能代表真实在线表现。本文建议把路由错误作为独立失败类别：系统可能拥有足够证据，却选择了不适合该问题的组织方式。定位这一点有助于避免无必要地重做整个图谱。

A multi-hop question connects several evidential steps, but the number of steps is not identical to reasoning difficulty. A two-step question may require difficult entity disambiguation, whereas a longer

path may involve an explicit hierarchy. The system should identify the required relationships, object types, and conditions rather than mechanically increase traversal depth. Asking which counters serve dishes containing chicken requires connections among counters, dishes, and ingredients. Adding a date also requires the relevant menu version. If one condition is missing, a connected path is still an incomplete answer.

Query routing can assign a request to factual lookup, relational expansion, global summarization, cross-modal verification, or a combination of these operations. The router can itself fail, so fallback behavior and execution records should be preserved. A useful study compares fixed routing, rule-based routing, and model-based routing while holding the corpus and generation budget constant. An advantage obtained by using test answers or by manually selecting the best route after seeing the outcome does not represent online performance. We recommend classifying routing mistakes separately. A system can possess adequate evidence but choose an unsuitable way of organizing it for the question. Recognizing that failure prevents unnecessary reconstruction of the whole graph. It also supports a more precise experiment: improve route selection while keeping extraction and generation unchanged, then determine whether the affected question class improves without an unacceptable loss elsewhere.

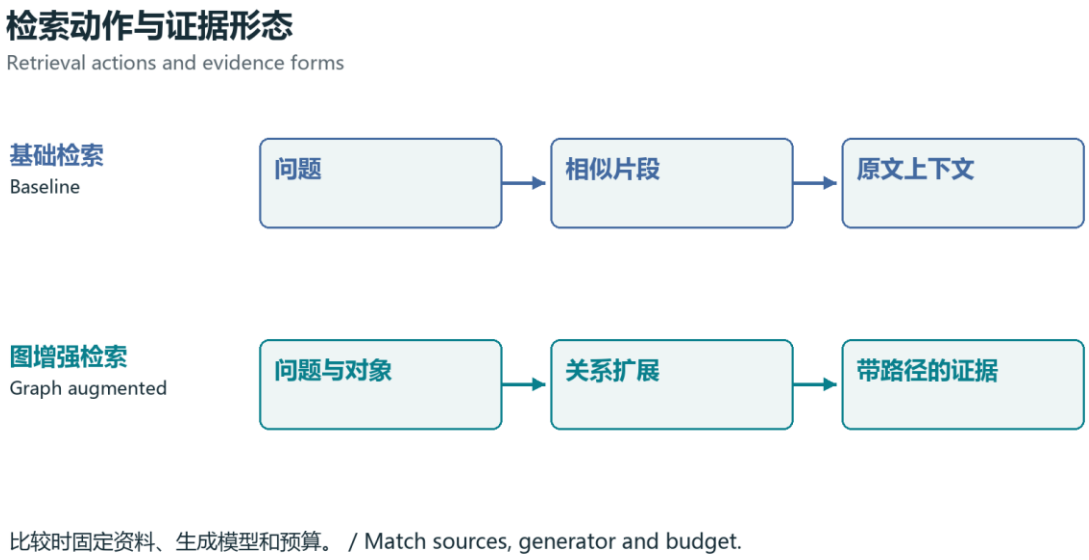


图 03 不同检索路径应按任务比较 / Compare retrieval routes by task
来源 / Source 本报告原创设计 / Original design

端到端失败的分解

Decomposing End to End Failure

一个错误答案可能来自数据、抽取、检索、组织、生成或展示环节。原文缺少必要信息时，再强的检索器也无法提供真实依据；实体错误合并时，图检索可能忠实返回错误路径；证据正确却未进入最终上下文时，问题属于组织或预算；模型忽略清晰证据时，则需检查生成行为。只有端到端分数，很难区分这些原因。评价应保留各阶段的中间对象，形成从结果回查输入的诊断链。[R25][R26]

本文提出按最早可定位失败点归因的实践方法。首先检查数据是否包含足够证据，再检查表示是否保留证据，然后检查召回和排序，最后检查结论及引用。对同一问题可构造“提供正确证据”的对照条件，以观察生成层在理想检索下是否仍然失败。这个对照不代表部署性能，但有助于测量改进检索的上限。错误分析还应包含正确答案使用错误依据的情况，因为只检查答案文本可能把这种脆弱成功漏掉。可靠系统需要同时解释回答是否正确以及为什么有资格这样回答。

An incorrect answer can originate in the data, extraction, retrieval, organization, generation, or presentation stage. If the source lacks the required information, even an excellent retriever cannot supply genuine support. If entities were incorrectly merged, graph retrieval may accurately return an invalid path. If correct evidence was retrieved but excluded from the final context, organization or budgeting is responsible. If the model disregarded clear evidence, generation behavior requires examination. An end-to-end score alone cannot distinguish these causes. Evaluation should retain intermediate objects at each stage so that an outcome can be traced back to its inputs. [R25][R26]

We propose attributing errors to the earliest failure that can be located. First check whether the collection contains sufficient evidence, then whether the representation preserves it, then inspect candidate retrieval and ranking, and finally examine the conclusion and its citations. A condition that supplies the correct evidence can reveal whether the generator still fails under ideal retrieval. This condition is diagnostic rather than a measure of deployment performance, but it helps estimate the potential benefit of improving retrieval. Error analysis should also include answers that happen to be correct while relying on the wrong evidence. Text-only scoring can overlook such fragile successes. A dependable system must address both whether the answer is correct and why the available material authorizes that answer. This decomposition makes improvement more economical, because each intervention can target a demonstrated failure rather than changing several components at once and obscuring the cause of any gain.

返回目录 Back to contents

第4章　微软 GraphRAG 的工作机制

Working Mechanisms of Microsoft GraphRAG

图索引的处理对象

Objects Managed by the Graph Index

微软 GraphRAG 的官方数据流程区分文档、文本单元、实体、关系、社区和社区报告等对象。这样的区分使检索系统同时保留原始材料与派生结构。图索引不是把整份文档简单改成几个关键词，而是把可定位片段与抽取对象连接起来，再组织更大范围的语义结构。报告中的图 5 将原始层、结构层与摘要层分开，帮助识别每一层增加了什么，又可能丢失什么。[R32]

本文认为，设计图索引时应首先保证对象之间的追溯关系。实体的名称可以变化，文本单元的分割可以调整，但旧回答所用的证据应仍能被解释。索引更新如果只覆盖当前图而不保留版本，就可能使历史回答无法复核。对研究实验而言，固定语料快照、切分方案和抽取版本同样重要，否

则比较中的“同一资料集”可能已经发生变化。图索引的评估因此既要检查查询质量，也要检查输入输出之间的身份稳定性。后者通常不直接反映在问答分数中，却决定系统能否持续维护。

The official Microsoft GraphRAG dataflow distinguishes documents, text units, entities, relationships, communities, and community reports. These categories allow original material and derived structure to coexist. A graph index does more than replace a document with a few keywords: it links locatable passages to extracted objects and then organizes larger semantic structures. Figure 5 separates the source, structural, and summary layers so that the information added or potentially lost at each transformation can be examined. [R32]

In our analysis, preserving traceability among these objects is the first design requirement. An entity's name may change, and a document may later be segmented differently, but the evidence used by an earlier answer should remain interpretable. If an update overwrites the current graph without preserving versions, historical answers can become impossible to audit. A research experiment likewise needs a fixed corpus snapshot, segmentation scheme, and extraction version. Otherwise the supposedly identical collection used in a comparison may have changed. Evaluation of a graph index should therefore consider both query quality and the stability of identity across transformations. The latter rarely appears directly in a question-answering score, yet it determines whether the system can be maintained over time. A representation that produces an attractive demonstration but cannot explain which source version supports a relationship is incomplete for applications requiring accountability or reproducible analysis.

实体关系抽取与信息压缩

Entity Relationship Extraction and Information Compression

大模型抽取为非结构化资料提供了灵活入口，但它同时是一种有损转换。原文中的条件、语气、例外和时间可能在短关系中被压缩。把“某报告推测项目可能延迟”写成“项目发生延迟”，会把来源、推测状态与可能性同时删除。本文建议让抽取记录保留命题状态和支持片段，并针对否定、条件句和转述构造测试。抽取结果不宜因符合输出格式，就被视为语义正确。[R02][R32]

抽取粒度应服务于查询任务。如果只保存高频实体，罕见但决定答案的对象可能丢失；如果把几乎每个名词都建成实体，图会变得嘈杂且维护困难。研究可设置不同的实体类型范围和关系约束，观察它们对召回、错误传播与索引成本的影响。生成模型还可能把一般知识补入抽取结果，因此必须检查每条关键关系是否来自当前资料。本文主张把模型提出的候选与通过证据检查的关系分开存储，使抽取成为可审计的知识构建过程，而不是一次不可逆的文本改写。

Language-model extraction provides a flexible entry point for unstructured material, but it is also a lossy transformation. Conditions, modality, exceptions, and time can disappear when a passage is compressed into a short relationship. Converting a report suggesting that a project might be delayed into a statement that the project was delayed removes the source, the inferential status, and the uncertainty at once. We recommend preserving the status of a proposition and its supporting passage, and testing extraction with negation, conditional statements, and reported speech. Compliance with an output schema does not establish semantic correctness. [R02][R32]

Extraction granularity should serve the query task. Retaining only frequent entities may remove a rare object that determines an answer. Turning almost every noun into an entity can instead produce a noisy

graph that is costly to maintain. A study can vary the permitted entity types and relationship constraints, then examine their effects on recall, error propagation, and indexing cost. A generator may also insert general knowledge into an extraction output, so important relationships should be checked for support in the current collection. We argue for storing model-proposed candidates separately from relationships that have passed an evidence check. Extraction then becomes an auditable process of knowledge construction rather than an irreversible rewriting of text. This distinction enables targeted correction and lets an application use more permissive candidates for exploratory retrieval without representing them as confirmed facts in a final answer.

社区层级与主题表征

Community Hierarchies and Theme Representation

社区检测依据图连接组织节点，形成可供汇总的群组。Leiden 研究讨论了社区连通性等图结构性质，微软 GraphRAG 则利用社区及其层级组织摘要。需要注意，图结构上的良好社区并不自动等于业务上正确的主题分类。社区名称和报告仍需要语言解释，其质量受抽取图及摘要过程共同影响。图的分区、主题的命名和主题对用户是否有用，是三个应分别检查的问题。[R27][R32]

本文建议在研究中同时保留社区成员、层级关系与摘要依据。调节社区粒度可能让报告更概括，也可能把重要差异合并掉。若一个主题只在少数片段中出现，它可能不会成为显著社区，却仍对某个用户极其重要。因此，全局理解不能简单等同于统计上最大的几个群组。对用户问题应检查主题覆盖与少数但相关的证据，并说明报告依据的是哪些层级。社区结构适合作为组织资料的工具，其输出仍需要与原始材料和使用目的对照，而不是被当作资料天然存在的唯一分类。

Community detection organizes nodes according to graph connections, producing groups that can be summarized. The Leiden study addresses structural properties such as community connectivity, while Microsoft GraphRAG uses communities and their hierarchy to organize summaries. A structurally well-formed community is not automatically a correct business topic. Community names and reports require linguistic interpretation, and their quality depends jointly on extraction and summarization. The graph partition, the naming of a theme, and the usefulness of that theme to a user are separate evaluation questions. [R27][R32]

We recommend retaining community membership, hierarchical relationships, and the evidence underlying each report. Changing granularity can produce a broader account or merge distinctions that matter. A topic appearing in only a few passages may not form a prominent community while remaining essential to a particular user's question. Global understanding should therefore not be equated with selecting the largest statistical groups. Evaluation should examine both topic coverage and small but relevant bodies of evidence, and state which hierarchy level supplied the reports. Community structure is a means of organizing material rather than a uniquely correct taxonomy inherent in the data. Different partitions may support different questions. Recording the chosen partition and its limitations makes the resulting analysis easier to interpret and avoids treating an algorithmic grouping as if it were an independently established organizational or causal relationship.

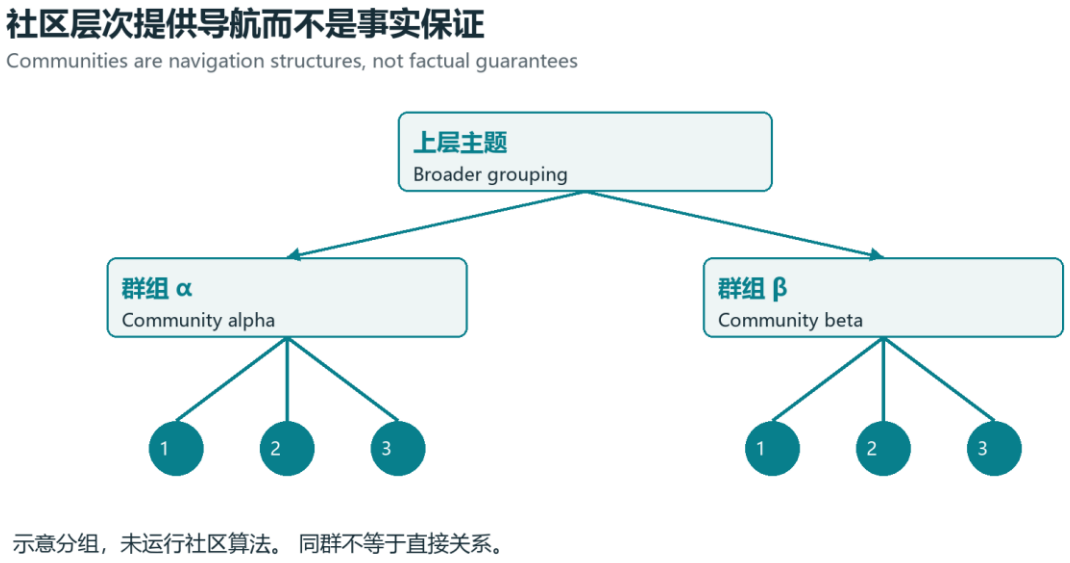


图 04 层次社区示意与解释边界 / Schematic hierarchy and interpretation boundary
来源 / Source 本报告教学示意，未执行聚类 / Schematic, no clustering run

全局检索与分批汇总

Global Search and Batched Synthesis

全局检索使用社区报告作为上下文，分批形成与问题相关的中间回答，再筛选和汇总为整体回答。这里有两个不同时间点的摘要：索引时生成的社区报告，以及查询时面向具体问题形成的中间回答。把两者混为一谈，会误以为系统预先知道了所有用户问题。官方流程说明了分批与聚合机制，但实际效果仍取决于报告是否保留了相关信息，以及聚合时是否正确处理重复和差异。[R33]

本文把全局汇总看成受预算约束的证据压缩。每一层压缩都可能丢失少数观点、例外或冲突，因此最终答案应避免把概括写成无条件的全体判断。研究可以追踪一个重要命题在原文、社区报告、中间回答和最终回答中的变化，检查语气是否逐步变强。不同批次的报告还可能重复引用同一来源，若把重复次数当作独立支持，就会高估证据强度。聚合阶段应识别来源重叠，并尽量保留分歧，而不是仅按语言流畅度合并。这样的检查比单纯观察答案长度更能解释全局摘要的可靠性。

Global search uses community reports as context, generates question-specific intermediate responses in batches, and filters and aggregates them into a final response. This involves summaries created at two different times: community reports produced during indexing and intermediate answers generated for a particular query. Confusing them can create the impression that the system anticipates every future question. The official workflow explains batching and aggregation, but their effectiveness still depends on whether reports retain relevant information and whether aggregation handles repetition and differences appropriately. [R33]

We analyze global synthesis as evidence compression under a resource budget. Each layer can lose minority views, exceptions, or conflicting statements. The final answer should therefore avoid converting a summary into an unconditional claim about every item in the collection. A study can follow an

important proposition through the source passage, community report, intermediate answer, and final response, examining whether its degree of certainty grows during transformation. Reports from different batches may also cite the same original source. Counting their repetition as independent corroboration overstates evidential strength. Aggregation should identify overlapping provenance and preserve meaningful disagreement rather than merge content solely for fluency. These checks explain the reliability of global summarization more effectively than answer length alone. They also make it possible to locate a failure at the layer where a qualification was lost, rather than blaming the final generator for an omission that occurred earlier.



图 05 索引摘要与面向问题的汇总不是同一步 / Index reports and query synthesis differ
来源 / Source 本报告原创设计 / Original design

局部问题与全局问题的配合

Combining Local and Global Questions

围绕具体实体的问题，通常需要细粒度关系和直接证据；围绕整个资料集合的问题，则更需要覆盖不同群组的概括。一个应用可能在同一会话中交替出现两类请求，例如先问某个项目的合作方，再问所有项目的共同风险。设计时应允许局部证据与较大范围摘要相互补充，并保持它们的来源层次。用户看见一个全局结论后，还应能追问它由哪些具体记录支持。

本文建议用问题的证据需求决定检索模式，而不是只根据问句长短或关键词判断。某个看似具体的问题可能需要跨文档比较，一个包含“总体”字样的问题也可能只涉及一小组明确记录。研究可由人工标注必要证据集合，再观察不同模式是否覆盖该集合。若同时运行多种模式，应记录成本并说明最终答案如何选取证据。把多个结果全部塞给模型，可能增加冲突和上下文噪声。模式组合的价值应体现为更合适的证据选择，而不是更多的检索调用或更长的回答。[R09][R33]

Questions centered on a particular entity generally need detailed relationships and direct evidence. Questions about a collection as a whole require a broader account spanning several groups. An application may alternate between these requests within one conversation, for example asking about one project's partners and then asking about risks common to all projects. The design should allow local evidence and wider summaries to complement one another while preserving their distinct levels of provenance. A user who sees a global conclusion should be able to ask which specific records support it.

We recommend selecting retrieval modes according to evidential requirements rather than question length or a few keywords. An apparently specific question may require comparison across documents, while a request containing the word overall may concern only a small, identifiable set of records. Researchers can annotate the necessary evidence set and examine whether different modes cover it. If several modes run together, their costs and the rule for selecting evidence should be recorded. Passing every retrieved result to the generator may introduce conflicts and context noise rather than improve the answer. The value of a combined mode should be demonstrated through more appropriate evidence selection, not through a larger number of retrieval calls or a longer response. Comparative work motivates examining such combinations, but a routing or fusion policy still requires validation for the intended corpus and workload. [R09][R33]

原文发现与适用条件

Findings in the Anchor Article and Their Conditions

微软早期文章报告，GraphRAG 在全面性、视角多样性和支持人工核查等比较维度上表现较好，同时其对来源的忠实性与基线相近。这组结论不能改写成“图谱消除了幻觉”，也不能变成适用于所有任务的准确率提升比例。后续论文围绕全局理解问题提供了更系统的分析，但应与早期演示分开引用。研究结果的解释必须保留所用语料、问题生成方法、评价者和比较对象。[R01][R02]

本文认为，最有价值的启示是评价单位发生了变化：当问题要求理解整批材料时，只评价少数片段与问句的相似性是不充分的。但这并不削弱向量检索在明确事实查找中的作用。图构建会引入成本、延迟和错误来源，社区报告也需要更新。是否采用这种方法，应取决于用户是否频繁提出关联和整体理解问题，以及系统是否有能力维护派生证据。合理的结论是建立任务条件下的适用判断，而不是把一种研究方法升级为所有知识工作的默认答案。

The early Microsoft article reports favorable comparisons on comprehensiveness, diversity of viewpoints, and support for human inspection, while reporting faithfulness to the source at a level similar to the baseline. These findings should not be rewritten as elimination of hallucination or as a universal percentage increase in accuracy. The later paper offers a more systematic analysis of global sensemaking questions, but it should be cited separately from the initial demonstration. Interpretation must retain the corpus, question-generation procedure, evaluator, and comparison system involved. [R01][R02]

The most useful implication, in our view, is a change in the unit of evaluation. When a question concerns an entire collection, similarity between the question and a few retrieved passages is insufficient. This does not remove the value of vector retrieval for explicit factual lookup. Graph construction introduces cost, latency, and additional opportunities for error, and community reports require maintenance. Adoption should depend on how often users ask associative or collection-level questions and whether the system can maintain the derived evidence. The appropriate conclusion is conditional suitability rather than treating one research method as the default answer to every knowledge task. A credible evaluation should include questions where a simpler baseline is expected to work well, as well as questions designed to expose the limitations that graph organization is intended to address. Such balance makes the boundary of a method's benefit visible and prevents selective demonstrations from becoming an exaggerated capability claim.

返回目录 Back to contents

第5章　图增强检索方法的比较与选择

Comparing and Selecting Graph Augmented Retrieval Methods

按知识组织和检索动作比较方法

Compare Knowledge Organization and Retrieval Actions

方法名称不能直接充当采购清单。一个系统自称 GraphRAG，可能指从文本抽取实体图，也可能指在现有业务图上查找邻居，还可能指将图编码器接入语言模型。本报告建议先回答四个具体问题：索引时保存什么，查询时从哪里出发，沿什么关系扩展，最终给生成器什么证据。用这四项建立比较表，才能识别两个名称不同的实现是否实际上做了相近的工作，也能发现同一名称下不可互换的技术路径。两篇综述为理解这一多样性提供入口，但具体实现仍要回到相应论文和版本。[R07][R08]

比较单位应是完整配置，包括数据处理、模型、检索预算、提示词、缓存和访问范围，而不能只有算法标签。假设方案甲使用更强的生成模型、更多原文和更长上下文，即使回答更好，也不能直接把差异归因于图结构。实际选型可以先建立一个可靠的向量与关键词混合基线，再只改变证据组织方式，随后逐步开放其他优化。这种顺序让团队知道图带来的增量在哪里。若图没有改善目标问题，结论应该是当前图配置尚无价值，而不是数据天然不适合任何图方法。

A method name is not a procurement checklist. A system called GraphRAG may extract an entity graph from text, traverse an existing business graph, or connect a graph encoder to a language model. This report proposes four concrete comparison questions: what is stored during indexing, where a query starts, which relations support expansion, and what evidence reaches the generator. These questions reveal when differently named systems perform similar work and when implementations sharing a name follow incompatible technical routes. The two surveys provide entry points into this diversity, while implementation decisions still require the corresponding paper and version.[R07][R08]

The comparison unit should be a complete configuration covering preprocessing, models, retrieval budgets, prompts, caching, and access scope. Suppose system A uses a stronger generator, more source material, and a longer context. Better answers cannot then be attributed directly to its graph. A practical selection process first establishes a reliable vector and keyword hybrid baseline, changes only evidence organization, and subsequently introduces other improvements. This sequence identifies the additional value of the graph. If the graph does not improve the target questions, the supported conclusion is that the current configuration has not demonstrated value. It does not establish that the data can never benefit from any graph method.

LightRAG 与不同粒度的检索线索

LightRAG and Retrieval Clues at Different Granularities

LightRAG 将图结构与检索结合，并强调不同层次的信息检索。理解它时，一个实用视角是把查询拆成具体对象线索和较抽象的主题线索：前者帮助定位某个实体及其邻接关系，后者帮助寻找跨

对象的关联信息。这与微软社区报告路线的侧重点并不完全相同，因此不能把两者当作同一索引的简单快慢版本。论文标题中的轻量与快速也不应替代本地测量；文本长度、抽取模型、图密度和更新方式都会改变实际代价。[R10]

例如，企业问题“设备 A 为什么连续停机”既包含明确设备，也包含需要解释的故障主题。系统可以并行寻找设备记录、故障类型以及相关工单，再按时间和来源整合。这里最容易犯的错误是把主题相似当成同一事件：两个设备都出现过高温，并不意味着它们共享一次故障。为避免这种扩展失控，检索线索应附带事件、地点和时间边界。评估时分别统计对象命中、关系覆盖与无关扩展比例，并记录这些线索是否帮助找到了原先缺失的证据。算法提供线索组织方式，任务边界决定它是否真正有用。

LightRAG combines graph structure with retrieval and emphasizes information retrieval at different levels. A useful reading distinguishes clues about specific objects from more abstract thematic clues. The former help locate an entity and its neighborhood; the latter help find information connecting multiple objects. Its emphasis differs from the Microsoft community report route, so these approaches should not be treated simply as faster and slower versions of one index. The words simple and fast in a paper title also cannot replace local measurements. Text length, extraction models, graph density, and update procedures all affect actual cost.[R10]

Consider an enterprise question asking why equipment A repeatedly stopped. It names an identifiable machine while also requiring an explanation of a fault pattern. A system might retrieve equipment records, fault categories, and associated work orders in parallel, then organize them by time and source. A common mistake is treating thematic similarity as event identity. Two machines experiencing high temperatures do not necessarily share one incident. Retrieval clues therefore need event, location, and temporal boundaries to prevent uncontrolled expansion. Evaluation should measure object identification, relationship coverage, and irrelevant expansion separately, and record whether the clues recover previously missing evidence. The algorithm supplies an organization of clues; the task boundaries determine whether that organization is useful.

HippoRAG 系列与关联式记忆

The HippoRAG Family and Associative Memory

HippoRAG 以受神经生物学启发的长期记忆为研究出发点，利用图和关联式检索组织外部知识；后续 HippoRAG 2 进一步讨论从 RAG 走向非参数持续学习的记忆系统。这里的记忆首先是一种工程组织方式，不意味着系统具有人的经历或生理机制。它的研究价值在于：一个问题可能无法靠单条文本的局部相似度找到所有材料，图上关联传播可以提供另一条发现路径。是否获得更好的长期知识使用能力，仍需在明确的更新与查询任务中验证。[R11][R12]

在实际设计中，记忆应同时包含保留与遗忘规则。加入一份新报告后，需要判断它补充了旧结论、推翻了旧结论，还是只描述了另一个时间片。如果只是不断累加边，系统可能越来越容易找到过时关联。持续学习的评估也不能只看静态测试集：应设置按时间到达的新知识、修改后的事实、被撤回的来源和重复表达，检查旧任务是否保持、新任务是否学会以及失效信息是否停止被使用。这样的测试将记忆能力转化为可观察行为。一个可审计的外部记忆库，理应能解释某条关联何时进入、由谁支持、何时失效。

HippoRAG starts from a research analogy with long term memory and uses graphs and associative retrieval to organize external knowledge. HippoRAG 2 extends the discussion toward memory systems for nonparametric continual learning. Memory here is primarily an engineering organization; it does not imply human experience or a biological mechanism. The important research idea is that local similarity between a question and a single passage may miss necessary material, while association across a graph offers another discovery route. Improved long term knowledge use must still be demonstrated through explicit update and query tasks.[R11][R12]

An operational memory design needs rules for retention and forgetting. When a new report arrives, the system must determine whether it supplements an earlier conclusion, contradicts it, or describes a different period. Continually adding edges can make outdated associations easier to retrieve. Evaluation of continual learning should therefore extend beyond a static test set. It should introduce new knowledge over time, corrected facts, withdrawn sources, and repeated descriptions, then check retention of earlier tasks, acquisition of new tasks, and cessation of invalid information use. These tests turn memory into observable behavior. An auditable external memory should explain when an association entered the system, which sources support it, and when it ceased to apply.

KG2RAG 与证据片段之间的组织

KG2RAG and the Organization of Evidence Chunks

KG2RAG 利用知识图谱引导片段扩展和组织，为理解“先召回再连起来”提供了有用研究实例。它提示我们，图的价值可以体现为改善证据上下文，而不一定要求生成器直接阅读一张大型图。一个段落谈设备，另一个段落谈维护记录，第三个段落说明替换部件；通过共享实体和关系，可以将这些片段组织成更连贯的回答材料。这里只概括论文的方法方向，不把其特定实验结果推广为所有数据上的优势。[R13]

组织证据时，需要保留逻辑连接的类型。两段材料共同提到一个供应商，属于共同实体；一份记录明确引用另一份文件，属于文档依赖；一个事件发生在另一个事件之后，属于时间顺序。它们对答案的支持力度不同。把所有连接压成无类型邻接，会让生成器把“相关”误读成“解释”。建议输出证据包时列出连接理由和原文定位，并限制跨越的最大步数。还应把直接证据放在推断路径之前，让读者先看到原文说了什么，再看到系统如何把材料连接起来。这样即使组织策略失败，也能定位到具体的错误连接。

KG2RAG uses a knowledge graph to guide chunk expansion and organization, offering a useful example of retrieving material and then connecting it. Its broader lesson is that a graph can improve the evidence context without requiring a generator to read one large graph directly. One passage may describe equipment, another maintenance records, and a third a replacement component. Shared entities and relations can organize them into more coherent answer material. This report summarizes that methodological direction without extending particular experimental findings into a universal advantage.[R13]

Evidence organization must preserve the type of logical connection. Two passages mentioning the same supplier share an entity. A record explicitly citing another file establishes a document dependency. One incident following another establishes a temporal order. These connections support an answer in different ways. Collapsing them into untyped adjacency encourages a generator to interpret relatedness as

explanation. An evidence package should therefore include the reason for each connection, source locations, and a maximum traversal depth. Direct evidence should appear before inferred paths, allowing readers to see what the sources actually state before examining how the system connects them. Even when the organization strategy fails, this arrangement makes the specific erroneous connection identifiable.

G-Retriever 与图本身成为问题对象

G-Retriever and Questions about the Graph Itself

有些任务不是从长篇文档建立图，而是已经存在带有文字属性的网络，用户直接询问其中的对象与联系。G-Retriever 研究文本属性图上的理解与问答，提醒我们不要把所有图增强生成都限定为微软式社区摘要。对于已有设备网络、论文引用网络或产品关系网，问题可能需要选取相关子图，再把结构与文字共同交给模型。该方向的训练与推理条件，应与仅调用外部语言模型的流水线分开说明。[R14]

选择子图不是越大越好。一个包含几千节点的连通区域，可能只含两条对问题真正必要的关系，却引入大量同名实体和外围描述。工程上可以把子图选择看作带预算的证据覆盖问题：优先保留能连接关键对象的路径、支持这些路径的来源以及必要的反例。若某条路径存在多种解释，应该输出备选子图而不是强迫模型选一个故事。图上的距离也不等于现实世界的重要性，一条很短但来自低可信来源的路径，可能不如更长但有明确记录的证据链。结构信息需要与证据质量共同判断。

Some tasks begin with an existing network whose nodes and edges carry text, and users ask directly about its objects and connections. G-Retriever studies understanding and question answering over textual graphs, reminding us that graph augmented generation is broader than Microsoft style community summaries. Existing equipment networks, citation networks, or product graphs may require selecting a relevant subgraph and presenting both structure and text to a model. Training and inference requirements for this route should be distinguished from pipelines that only call an external language model.[R14]

A larger selected subgraph is not automatically better. A connected region with thousands of nodes may contain only two relations necessary for a question, while introducing many ambiguous names and peripheral descriptions. Engineering can treat subgraph selection as evidence coverage under a budget: prioritize paths connecting key objects, sources supporting those paths, and necessary counterexamples. If a path permits several interpretations, return alternative subgraphs rather than forcing the model to choose one story. Graph distance is also not equivalent to real world importance. A short path from a weak source may be less useful than a longer chain with explicit records. Structural information must be judged together with evidence quality.

从任务矩阵形成组合方案

Build a Portfolio from a Task Matrix

本报告建议按任务而非品牌建立选型矩阵。明确事实查询优先检查精确定位与出处；跨文档关系问题关注实体消歧与路径完整性；全库主题问题关注覆盖、归纳和少数观点保留；视觉问题关注

图像证据是否不可替代。对每一行任务，同时记录最小可用方案、图带来的预期收益、可能增加的失败模式和可接受成本。比较研究已经强调 RAG 与 GraphRAG 的表现依赖任务条件，因此不能从单一平均分推出一种方法全面替代另一种。[R09]

最终系统可以组合多条检索路径，但组合必须有可解释的路由。先识别问题是否需要全局概括、关系推理或视觉辨认，再分配检索预算；若分类不确定，可同时运行较便宜的候选路径并合并证据。路由自身也要评估，因为一次错误的全局路由可能比检索模型的小幅提升影响更大。建议保留每次请求的路由理由、候选数量、证据来源和回退原因。采购或上线决策应基于真实问题分布下的效用，而不是最复杂问题上的演示效果。复杂方案值得采用的条件，是它持续改善重要任务并承担得起新增维护责任。

This report recommends a task matrix rather than a brand matrix. Explicit factual questions prioritize precise localization and provenance. Relationships across documents require entity disambiguation and complete paths. Questions about corpus wide themes require coverage, synthesis, and preservation of minority views. Visual questions require checking whether image evidence is indispensable. For every task row, record the minimum viable method, the expected benefit of a graph, additional failure modes, and acceptable cost. Comparative research emphasizes that RAG and GraphRAG performance depends on task conditions; one average score cannot establish complete replacement of one approach by another.[R09]

The final system may combine several retrieval routes, but their orchestration needs an interpretable router. Determine whether a question requires global synthesis, relational reasoning, or visual identification before allocating retrieval budgets. When classification is uncertain, inexpensive candidate routes can run together and their evidence can be merged. Routing itself needs evaluation: an incorrect global route may matter more than a small improvement in the retriever. Preserve routing reasons, candidate counts, evidence origins, and fallback explanations for each request. Procurement and deployment decisions should follow utility under the real distribution of questions, rather than demonstrations on unusually complex cases. Complexity is justified when it consistently improves important tasks and the organization can maintain its additional responsibilities.

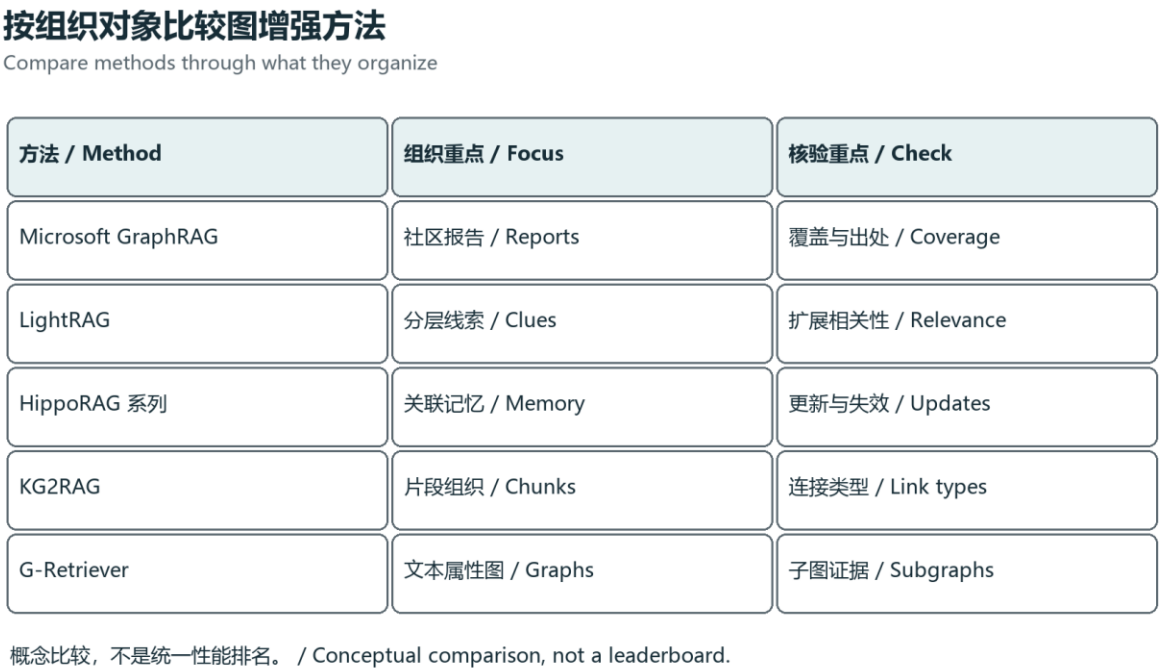


图 06 方法比较矩阵 / Method comparison matrix

来源 / Source 依据 R02、R10—R14 的方法方向综合 / Synthesis of method directions

返回目录 Back to contents

第6章 多模态知识图谱的表示与对齐

Representation and Alignment in Multimodal Knowledge Graphs

模态成为属性节点还是观察证据

Modalities as Attributes Nodes or Observational Evidence

多模态知识图谱把不同形式的信息纳入同一知识组织，但“纳入”存在不同深度。最简单的方式是在实体上附一张图片；进一步可以把图片作为独立节点，保存拍摄时间、来源与许可；更细的方式会把图片区域、文字框、音频区间或视频片段建成可定位的证据对象。三种方式各有用途：展示型目录可能只需属性，跨文档识别需要独立资源，细粒度问答通常需要局部证据。多模态图谱综述讨论了不同构建与学习任务，不能把有图片附件的数据库自动视为完成了语义对齐。[R15][R16]

设计应从需要回答的问题倒推。如果用户问“这道菜的配料表在哪里”，图片资源和文档位置足够；如果问“图中左上方的绿色叶片是什么”，就需要区域坐标与辨认依据；如果问“视频中哪一步加入调味料”，则需要时间区间。粒度越细，标注、索引和维护代价越高。因此建议建立逐级细化策略：先保存完整原始资源，再按实际查询需要增加局部结构，避免一开始对所有图片进行无目标的密

集标注。局部结构还必须保留回到原图的变换关系，否则缩放裁剪后，正确的框也可能指向错误位置。

A multimodal knowledge graph organizes information in different forms, but inclusion can happen at several depths. The simplest approach attaches an image to an entity. A richer approach makes the image an independent node with capture time, provenance, and licensing information. A finer representation makes image regions, text boxes, audio intervals, or video segments into locatable evidence objects. Each has a role: a display catalogue may need only attributes, identification across documents requires independent resources, and fine grained questions usually require local evidence. Surveys cover different construction and learning tasks; attaching pictures to a database does not by itself establish semantic alignment.[R15][R16]

Design should work backward from questions. Asking where a dish ingredient list appears requires a resource and document location. Asking what the green leaves in the upper left of an image are requires region coordinates and identification evidence. Asking when seasoning enters a cooking video requires a time interval. Finer granularity raises annotation, indexing, and maintenance costs. A progressive strategy therefore preserves complete original resources first and adds local structure when queries justify it, instead of densely annotating every image without a defined purpose. Local structures must also preserve transformations back to the original resource. Otherwise, a correct bounding box can point to the wrong location after resizing or cropping.



图 07 多模态资源到细粒度证据的层次 / Granularity from resources to evidence
来源 / Source 本报告原创设计 / Original design

共享嵌入空间提供相似线索

Shared Embedding Spaces Supply Similarity Clues

CLIP 展示了利用图文配对监督学习可迁移视觉表示的路径，为跨模态检索提供重要基础。共享空间让文本查询和图片候选可以计算相似度，却没有自动赋予相似度以事实语义。文字“宫保鸡丁”和一张菜品照片相近，可能因为颜色、摆盘或训练中的共现模式；它不能独立证明这张照片对应某个门店的真实商品，更不能证明其中不存在不可见原料。嵌入分数首先是候选排序信号，应与商品标识、上下文和来源信息一起使用。[R17]

应用中最好把相似、匹配与确认设置为不同状态。相似表示模型建议查看；匹配表示通过一组业务条件；确认表示已达到明确审核规则。每次状态变化记录依据，能避免将视觉模型输出悄悄升级为知识事实。对难以辨认的细节，应允许返回多个候选或不知道，并把需要补充的证据说明清楚。评估也不只计算总体检索命中，还要加入外观相近但身份不同的困难负例，例如同名菜在不同门店的版本、相同包装中的不同产品，以及背景很像却属于另一对象的照片。这样的负例能揭示模型依赖了什么捷径。

CLIP demonstrates transferable visual representation learning from paired images and text and provides an important foundation for cross modal retrieval. A shared space makes similarity between a text query and image candidates computable, but it does not automatically give that similarity factual meaning. A picture close to the phrase kung pao chicken may match through colour, presentation, or training cooccurrence. It does not independently establish that the image shows a particular restaurant product, and it cannot establish the absence of invisible ingredients. Embedding scores should first be treated as candidate ranking signals and combined with product identifiers, context, and provenance.[R17]

Applications should distinguish similarity, matching, and confirmation. Similarity means a model recommends inspection. Matching means a candidate passes specified business conditions. Confirmation means it satisfies an explicit review rule. Recording evidence for each transition prevents visual outputs from silently becoming knowledge facts. Ambiguous details should permit multiple candidates or an unknown answer, accompanied by an explanation of the missing evidence. Evaluation should include more than overall retrieval success. Hard negatives include visually similar but distinct items, different restaurant versions of a dish with the same name, different products in similar packaging, and photographs with matching backgrounds but different subjects. Such cases expose the shortcuts on which a model relies.



图 08 相似、匹配与确认的状态区分 / Similarity, matching and confirmation
来源 / Source 本报告原创设计 / Original design

实体链接是跨模态身份判断

Entity Linking Is a Cross Modal Identity Decision

实体链接要回答某个提及、图像区域或商品记录指向哪一个已存在实体。它不同于识别类别：认出一盘鸡肉菜属于视觉分类，把它链接到某门店某日期的菜单商品则需要身份条件。多模态实体链接研究尝试结合视觉、文字和图结构线索，但来源稀缺、同名和域变化仍应在具体数据中检验。本报告将 KGMEL 作为这一研究方向的检索入口，不依据摘要之外的材料推断其全部实现细节或部署性能。[R22]

可实施的链接流程应包含候选生成、条件核验和拒绝机制。先用别名、编号和向量召回候选，再比较门店、时间、包装文字及邻接关系。若两个候选都合理，不应强制合并，而应建立待判定的观察节点。还可以把同一实体与同一类别作为不同关系保存，让系统表达“看起来属于这类菜，但无法确认是哪份商品”。误合并通常比漏链接更难修复，因为它会把不同对象的证据混在一起并影响下游摘要。对于高影响实体，应优先降低误合并率，同时提供人工快速检查界面。正确拒绝也是链接系统的有效输出。

Entity linking asks which existing entity a mention, image region, or product record refers to. This differs from recognizing a category. Identifying a chicken dish is classification; linking it to a restaurant menu item on a particular date requires identity conditions. Multimodal entity linking research combines visual, textual, and graph clues, but sparse evidence, ambiguous names, and domain changes require testing on the actual data. This report uses KGMEL as an entry point into that direction without inferring complete implementation details or deployment performance from material beyond the inspected abstract.[R22]

A practical linking process needs candidate generation, condition checking, and rejection. Aliases, identifiers, and vectors first retrieve candidates. Restaurant, time, packaging text, and neighboring relations then help distinguish them. If two candidates remain plausible, preserve an unresolved observation instead of forcing a merge. Separate relations for same entity and same category allow the system to say that an image appears to show a category of dish while the specific product remains unidentified. Incorrect merges are often harder to repair than missed links because they mix evidence from different objects and contaminate downstream summaries. High impact entities therefore justify prioritizing a low false merge rate and providing efficient human inspection. Correct rejection is a valid linking outcome.

局部证据与版面信息不可随意压平

Preserve Local Evidence and Document Layout

文档中的图表和版面本身携带关系。一个数值位于哪个列标题下，图片注释对应哪个对象，流程箭头连接哪两个步骤，都可能在纯文本转换时丢失。ColPali 以视觉语言模型开展文档检索研究，提供了直接利用文档视觉表征的另一思路；它与显式知识图谱是可组合而非等同的技术。保留页面图像可以弥补文字解析的损失，但若要解释答案，仍需指出具体页、区域和相关文字。[R23]

对于研究报告或设备手册，建议同时保留逻辑结构和视觉坐标。表格单元格应能回到行列标题，图中标记应能回到图例，跨页续表应保持表身份。生成答案时可用文字说明结论，并附局部图像

证据供读者核验。这里需要防止一种新的错误：裁剪图像去掉单位、坐标轴或否定注释后，看似清楚的局部证据反而改变了原意。因此局部证据应包含必要上下文，并能一键查看整页。图像是否有助于答案，可以通过保留与移除图像的对照测试判断；仅仅让回答页面更漂亮，不构成多模态检索的实质收益。

Figures and document layout carry relations. Which column header governs a number, which object a caption describes, and which steps an arrow connects may all disappear during plain text conversion. ColPali studies document retrieval using vision language representations, offering a route that uses visual document information directly. This approach can complement an explicit knowledge graph but is not equivalent to one. Preserving page images may recover information lost by text parsing; explaining an answer still requires identifying the relevant page, region, and text.[R23]

Research reports and equipment manuals should preserve both logical structure and visual coordinates. Table cells should link back to row and column headers, figure labels to legends, and continued tables to a stable table identity. Answers can state a conclusion in text and attach a local image for verification. A new failure needs attention: cropping away units, axes, or a qualifying note may make a local image look clearer while changing its meaning. Local evidence should therefore include necessary context and provide access to the full page. Image contribution can be tested by comparing conditions with and without images. A more attractive answer page alone does not demonstrate a substantive benefit from multimodal retrieval.

缺失冲突与不可靠模态

Missing Conflicting and Unreliable Modalities

多模态融合不应假设每种模态都完整而可靠。商品可能只有文字没有图片，照片可能过期，视频声音可能与画面不同步，评论可能谈的是过去的版本。如果模型把缺失模态当作零值向量，很容易学到与真实内容无关的模式；如果强迫不同模态一致，则可能掩盖数据中的真实冲突。建议为每个观察保存可用性、质量、时间和适用范围，并在检索时显式使用这些信息。某一种模态缺失，应该影响系统能回答什么，而不是由模型默默补齐事实。

冲突需要先定位层次。文字说有花生、图片未检出花生，不一定矛盾，因为花生可能不可见；文字明确说不含花生、另一份同版本配料记录说含花生，才构成需要处理的陈述冲突。系统可以区分内容矛盾、身份不确定、时间不一致和观察能力有限四种原因，并给出相应的补充证据请求。训练与测试中应主动遮蔽某些模态、替换过期图片、加入错误配对，检查系统是否能降低确信程度。融合的成熟标志，不是所有输入都被合成一个流畅答案，而是系统知道哪些输入不能共同支持该答案。

Multimodal fusion should not assume that every modality is complete and reliable. A product may have text but no image, a photograph may be outdated, audio may be misaligned with video, and a review may describe an earlier version. Treating missing modalities as ordinary zero vectors can encourage patterns unrelated to content. Forcing agreement across modalities can conceal real conflicts. Each observation should therefore retain availability, quality, time, and scope, and retrieval should use these explicitly. Missing information should change what the system can answer instead of inviting the model to silently fill factual gaps.

Conflict diagnosis must identify the relevant level. Text stating that peanuts are present and an image detector finding none are not necessarily contradictory because peanuts may be invisible. An explicit claim of no peanuts and an ingredient record asserting peanuts for the same version create a statement conflict. Systems can distinguish contradictory content, uncertain identity, temporal mismatch, and limited observability, then request appropriate additional evidence. Training and testing should deliberately mask modalities, substitute outdated images, and introduce incorrect pairings to check whether confidence decreases appropriately. Mature fusion is not the ability to turn every input into one fluent answer. It is the ability to recognize which inputs cannot jointly support that answer.

从多模态图谱走向多模态 GraphRAG

From Multimodal Graphs to Multimodal GraphRAG

多模态图谱主要解决知识如何表示、关联和更新，多模态 GraphRAG 还要解决如何围绕问题选择证据并生成可核验回答。MuRAG 提供图像与文字外部记忆的早期研究实例；MMGraphRAG、mKG-RAG 以及 RAG-Anything 则体现了后续将多模态信息与结构化组织结合的不同探索。这些名称属于不同研究系统，不能拼接成一个已经标准化的统一架构。本报告只把它们作为设计空间的证据，不在缺少同条件复现时排列统一性能榜单。[R18][R19][R20][R24]

可迁移的设计原则是让图谱、检索和生成共享同一证据标识。图像区域支持哪个观察，观察链接哪个实体，检索为何选择它，答案哪句话依赖它，都应能串起来。若图谱只保存概念、检索只返回图片文件名、答案只给一段总结，这三部分虽然都运行成功，却无法共同完成证据追踪。建议将最小输出定义为“回答陈述＋证据定位＋关系路径＋不确定性说明”。对不需要视觉信息的问题，不必强行增加视觉处理；对必须看图的问题，也不能仅用图片的自动描述代替全部原始证据。模态选择应由问题的证据需求决定。

A multimodal graph addresses representation, association, and updating. Multimodal GraphRAG must additionally select evidence for a question and generate a verifiable answer. MuRAG offers an early example of external memory containing images and text. MMGraphRAG, mKG-RAG, and RAG-Anything illustrate later attempts to combine multimodal information with structured organization. These names refer to different research systems, not interchangeable components of a standardized architecture. They are treated here as evidence of a design space rather than entries in one performance ranking without matched reproduction.[R18][R19][R20][R24]

A transferable principle is to share evidence identifiers across the graph, retrieval, and generation. An image region should connect to the observation it supports, the entity linked by that observation, the reason it was retrieved, and the answer sentence relying on it. If the graph stores only concepts, retrieval returns only image filenames, and generation produces only a summary, all three components may run successfully without jointly supporting traceability. A minimum output can instead contain an answer claim, evidence location, relation path, and uncertainty statement. Questions that do not need visual information should not incur visual processing automatically. Questions that require images should not depend entirely on automatically generated captions. Evidence needs should determine modality selection.

返回目录 Back to contents

第7章　美团菜品图谱与 CMIngre 案例分析

The Meituan Dish Graph and CMIngre Case Study

从菜名理解走向食材层次

From Dish Names to Ingredient Level Understanding

美团文章把研究重点放在食材级理解，并介绍了 CMIngre 基准以及菜品知识图谱业务应用。文章发表于 2024 年 5 月 17 日，对应论文题为 Toward Chinese Food Understanding: A Cross-Modal Ingredient-Level Benchmark。论文在线发表年份与正式卷册年份存在区别，报告在书目中分别记录，避免把一个项目误当成两个独立研究。文章本身更适合说明业务动机与流程，论文和数据集则用于追查基准定义与实验设置。[R03][R04][R05]

从研究问题看，菜名是一个压缩程度很高的标签。同一种名称可能对应不同原料、制作方法和地域版本，而相近原料也可能被加工成外观差异很大的成品。因此，单纯建立“菜名到类别”的映射很难支持精细商品理解。本报告据此提出一个阅读角度：把菜品名称看作入口，把食材、做法、口味与商家版本看作需要独立核验的属性。这里的属性关系不是由名称自动推出，而是由菜单文字、图片观察或其他记录支持。这个区别决定了图谱是在重复常识，还是在描述某个真实商品。研究案例的启发首先在于问题粒度的下沉。

The Meituan article focuses on ingredient level understanding and introduces the CMIngre benchmark alongside a business application for dish knowledge graphs. Published on May 17, 2024, it relates to the paper Toward Chinese Food Understanding: A Cross-Modal Ingredient-Level Benchmark. Online publication and journal volume dates differ, so the bibliography records them separately rather than treating one project as two independent studies. The article explains business motivation and workflow, while the paper and dataset support examination of benchmark definitions and experimental settings.[R03][R04][R05]

A dish name is a highly compressed label. The same name can refer to different ingredients, cooking methods, and regional versions, while similar ingredients can become visually different dishes. A mapping from name to category alone therefore offers limited support for detailed product understanding. This report proposes reading the case with names as entry points and ingredients, preparation, taste, and restaurant versions as separately verifiable attributes. These relations should be supported by menu text, visual observations, or other records rather than inferred automatically from the name. The distinction determines whether a graph repeats general expectations or describes a specific product. The first lesson of the case is the value of moving the research question to a finer level.

数据规模来源与标签层次

Dataset Scale Sources and Label Hierarchy

CMIngre 包含 8,001 组图像与食材信息，429 种食材标签以及 95,290 个边界框。图像来源分为菜肴 1,719 组、食谱 2,330 组和用户生成内容 3,952 组。食材检测实验另将细标签映射到 67 个较粗类别；

这两个类别数属于不同层级，不能互相替换。数据清洗还包含低频标签处理和语义合并，因此最终标签空间既反映领域概念，也反映可获得数据的条件。[R03][R05]

本报告的分析重点是这些数字背后的评估边界。三种来源可能具有不同清晰度、拍摄角度和文字完整性；如果随机混合划分训练测试，同一来源的共同风格可能让结果看起来比真实迁移更容易。细标签映射到粗标签可以降低辨认难度，但也会隐藏细分类别上的错误。对教学与应用而言，应同时报告标签层级、每类样本分布和来源划分。若需要复现实验，应固定数据版本与划分清单，并检查近重复图像。仅报告总样本量不足以说明困难程度，尤其当大量边界框来自少数常见食材时，平均指标可能掩盖长尾类别的弱点。

CMIngre contains 8,001 image and ingredient pairs, 429 ingredient labels, and 95,290 bounding boxes. Its sources comprise 1,719 dish pairs, 2,330 recipe pairs, and 3,952 user generated pairs. Ingredient detection experiments map fine labels to 67 coarser categories. These counts describe different levels and are not interchangeable. Cleaning includes treatment of infrequent labels and semantic merging, so the final label space reflects both domain concepts and available evidence.[R03][R05]

The analytical issue is the evaluation boundary behind these numbers. Sources can differ in clarity, camera angle, and textual completeness. A random mixed split may make transfer appear easier because training and testing share source styles. Coarsening labels reduces recognition difficulty while potentially hiding fine category errors. Educational and operational use should therefore report label level, per category distributions, and source splits together. Reproduction requires fixed dataset versions and split manifests, including checks for near duplicate images. Total sample size alone does not characterize difficulty, especially when many bounding boxes belong to a few common ingredients. Aggregate scores can conceal weaknesses on rare categories.

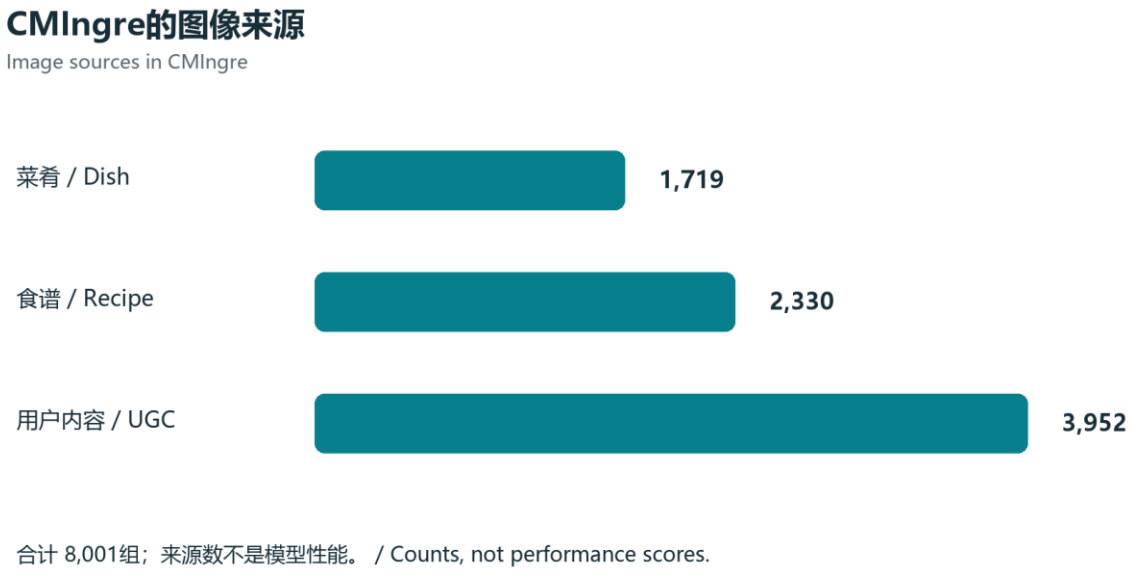


图 09 CMIngre 来源数量，单位为图像与食材对 / CMIngre pairs by source

来源 / Source R03、R05；本报告依原始数值重绘 / Redrawn from source counts



图 10 数据集细标签数与检测实验类别数需分开 / Fine dataset labels and coarser detection categories
来源 / Source 本报告原创设计 / Original design

检测与跨模态检索是两类任务

Detection and Cross Modal Retrieval Are Different Tasks

食材检测关注图片中的类别与位置，跨模态检索关注图片与食材组合的匹配排名。文章比较了检测基线，并讨论利用区域特征和自适应池化进行检索的两阶段方法。理解实验时，应把检测平均精度、检索召回与排名指标分开；检测框更准可能改善检索，但二者没有简单的一一对应关系。本报告不把文章中某种配置的比较结论外推为所有视觉模型的固定排序。[R03]

以一张食材混合程度很高的图片为例，检测器可能找到大块肉类却漏掉少量配料，检索器仍可能因为整体组合相近而找对菜谱。反过来，检测框位置不错，但食材类别混淆也可能把检索引向错误候选。改进实验应分别改变区域定位、区域分类和融合策略，再观察下游变化。若只比较最终检索分数，就难以判断改进来自哪个环节。还要检查查询方向：从图找食材与从食材找图的候选集合和歧义不同。对包含多种食材的查询，应说明相关性要求是全部匹配、部分匹配还是层次兼容。任务定义清楚，模型表现才有可解释意义。

Ingredient detection concerns categories and locations within images. Cross modal retrieval concerns the ranking of matches between images and ingredient combinations. The article compares detection baselines and discusses a two stage retrieval approach using region features and adaptive pooling. Detection average precision, retrieval recall, and ranking measures must be distinguished. Better boxes can help retrieval, but the relationship is not one to one. This report does not extrapolate comparisons between particular configurations into a permanent ranking of all visual models.[R03]

In an image containing mixed ingredients, a detector may identify large meat regions but miss small additions, while a retriever still finds a suitable recipe from the overall combination. Conversely, well positioned boxes with confused categories can direct retrieval toward the wrong candidates. Improvement studies should vary localization, classification, and fusion separately before examining downstream changes. Final retrieval scores alone cannot identify which component improved. Query direction also matters: retrieving ingredients from an image and retrieving images from ingredients have different candidate sets and ambiguities. For combinations, relevance must specify whether all ingredients, some

ingredients, or compatible hierarchy levels must match. Model performance becomes interpretable only after the task is defined clearly.

文本和图像如何进入菜品图谱

How Text and Images Enter the Dish Graph

文章的业务流程从文本抽取食材、口味、口感、菜系和烹饪方法，并用图像目标检测补充食材及区域信息，随后进行知识融合。这个流程与微软 GraphRAG 的主要连接点在于形成可检索的关系结构，而不是已经包含同一种社区摘要机制。美团案例首先展示构图与商品理解，微软案例主要讨论基于结构的发现与概括；把两者结合属于本报告提出的应用设计。[R03][R01]

组合时需要增加一个证据层。文本抽取结果应连接到具体文字片段，图像结果应连接到区域，二者只有在商品身份、版本与语义条件匹配后才能融合。假如评论说“上次没有放辣椒”，图片显示红色块状物，直接合并成“该菜含辣椒”会同时忽视时间和视觉歧义。建议在图谱中保留观察节点，记录模型候选和审核状态，然后再生成面向查询的属性视图。这样既能用多模态互补提高覆盖，也能保留模态之间的不一致。融合的目标是把证据组织得更可判断，而不是把所有输入压缩成没有出处的单一属性表。

The business workflow extracts ingredients, taste, texture, cuisine, and cooking methods from text, supplements ingredient and region information through image detection, and then fuses knowledge. Its connection to Microsoft GraphRAG lies in creating retrievable relationships, not in already implementing the same community summary mechanism. The Meituan case primarily demonstrates construction and product understanding, whereas the Microsoft case emphasizes discovery and synthesis using structure. Their combination is an application design proposed by this report.[R03][R01]

That combination needs an additional evidence layer. Text outputs should link to exact passages and visual outputs to regions. Fusion should occur only after product identity, version, and semantic conditions agree. If a review says that chilli was omitted last time while a photograph contains red objects, merging both into a permanent chilli attribute ignores temporal scope and visual ambiguity. Observation nodes should retain model candidates and review status before producing a query oriented attribute view. This supports complementary coverage without removing disagreement between modalities. Fusion should make evidence easier to judge rather than compress every input into an attribute table without provenance.

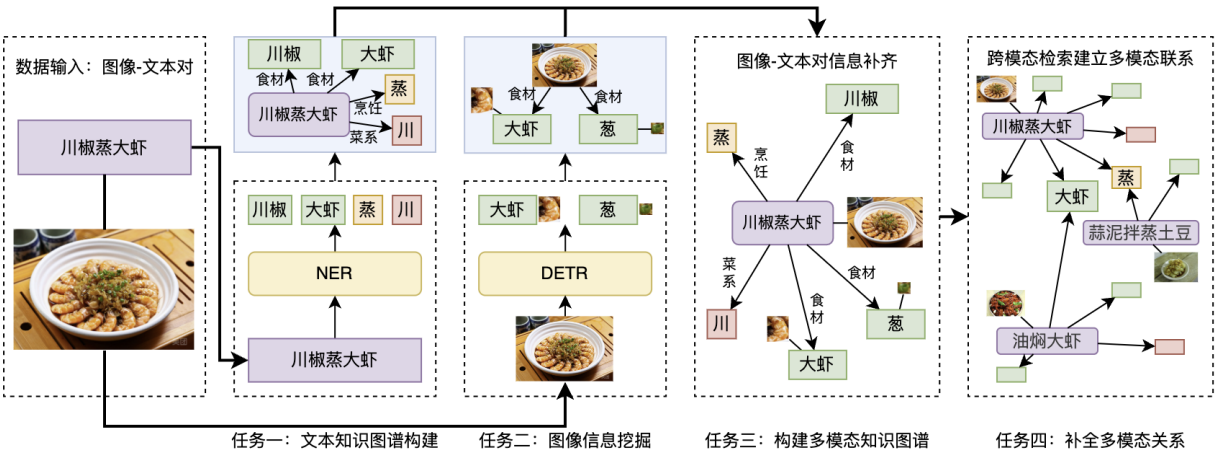


图 11 美团原文图 5.1 多模态菜品图谱构建流程 / Original Meituan Figure 5.1

来源 / Source R03 原图，文字 NER 与图像检测后进行融合；原图沿用作者标记 / Original source labels retained

业务数字的正确解释

Interpreting Business Numbers Correctly

文章报告，对比文本单模态，知识准确率由 95%到 96.52%，覆盖率由 80%到 87.01%；相同商品识别错误率由 20.38%到 2.3%。前两项分别增加 1.52 和 7.01 个百分点，最后一项下降 18.08 个百分点，相对下降约 88.71%。这些是文章所述项目评测结果，不是本报告独立复现，也不是 CMIngre 目标检测成绩。数字换算已用 Wolfram 及本地算术交叉检查。[R03]

解释这些结果时，需要知道评测单位、样本选择、审核标准和比较条件。若这些细节无法从已读取材料完整恢复，就不应擅自计算置信区间或声称收益可以按同样比例迁移到新门店。准确率与覆盖率也可能互相牵制：新增属性若主要来自更容易的商品，整体分数上升不能证明所有细分类别同步改进。业务错误率的改善可能包含图谱之外的系统因素，因此不能自动等同于某个视觉模型的因果贡献。新项目可借鉴的做法，是为自己的问题建立可复核基线、固定评测集并逐步消融，而不是复制一个百分比作为承诺。该案例提供可行性的实例，部署收益仍需要本地证据。

The article reports knowledge accuracy changing from 95% to 96.52%, coverage from 80% to 87.01%, and identical product recognition error from 20.38% to 2.3%. These correspond to increases of 1.52 and 7.01 percentage points and an error decrease of 18.08 percentage points, or approximately 88.71% relatively. They are reported project evaluations, not independently reproduced findings or CMIngre detection scores. Wolfram and local arithmetic were used to cross check the conversions.[R03]

Interpretation requires evaluation units, sample selection, review standards, and comparison conditions. Where inspected material does not establish these fully, confidence intervals should not be invented and transfer to new restaurants should not be promised at the same rate. Accuracy and coverage may also trade off. If newly covered attributes mainly concern easier products, an aggregate increase does not show improvement for every category. Business error changes can include system factors beyond the graph and cannot automatically establish the causal contribution of one visual model. New projects should borrow the discipline of a verifiable baseline, fixed evaluation set, and staged ablations rather than copy a

percentage as a commitment. The case demonstrates feasibility; deployment benefits still need local evidence.

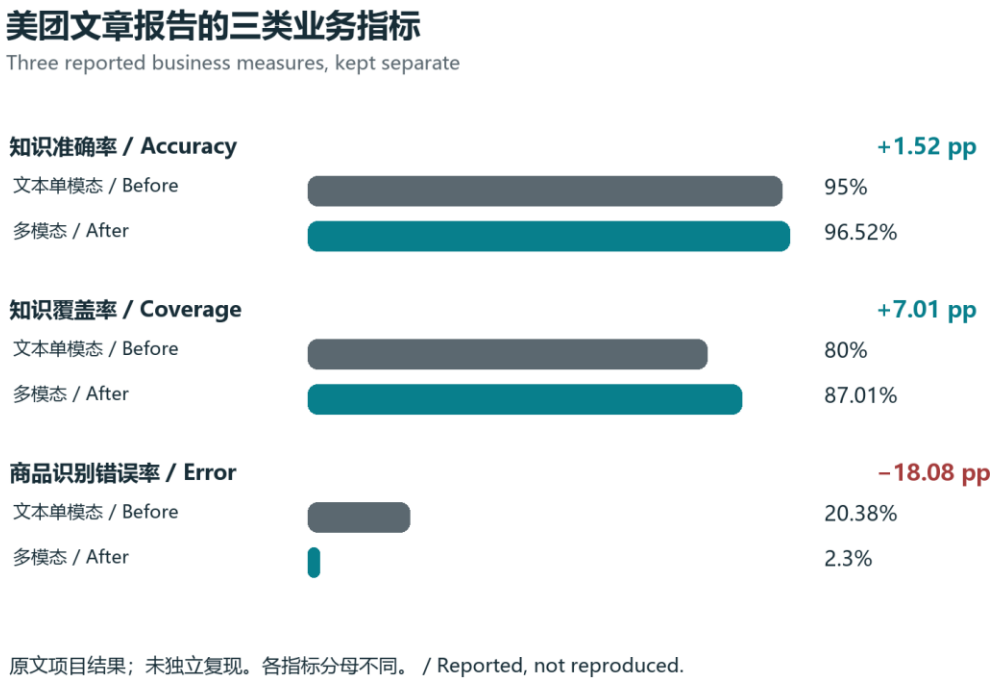


图 12 业务指标与百分点变化，不能混作检测成绩 / Reported business metrics and point changes
来源 / Source R03；Wolfram 与本地算术复算差值 / Arithmetic cross checked

可迁移的经验与不应迁移的假设

Transferable Lessons and Assumptions That Need Retesting

从美团案例可以提炼出三条设计经验：把业务对象细化到足以回答真实问题的粒度，利用多源输入补足单一来源的缺口，把构图能力放回实际业务任务中检验。本报告进一步强调，经验的迁移要经过数据和任务的重新定义。食材图谱中的层次关系可以启发零部件或文化物件分类，但不能直接复制类别体系；菜品图片中的局部检测可以启发工业视觉证据建模，但拍摄条件、可见性和错误代价完全不同。迁移的是问题解决结构，具体假设必须重测。

特别需要避免将视觉食材识别包装成过敏原或营养安全保证。不可见调料、加工过程和交叉接触不由照片完整表达，某道菜的名称也不提供可靠的排除证据。若系统面对这类需求，应返回已确认的配料来源、适用版本以及仍未知的部分，并指向可以核实信息的责任主体。报告后文的食堂设计只讨论信息组织与确认流程，不作临床或个体饮食判断。这个边界不是削弱图谱价值，反而使图谱能够承担合适的任务：帮助发现信息缺口、定位证据、减少误匹配，并把需要人工确认的事项清楚呈现出来。

Three design lessons can be drawn from the case: choose a business object granularity that answers real questions, use multiple sources to address gaps in a single source, and evaluate construction capabilities through actual tasks. Transfer still requires redefining data and tasks. Ingredient hierarchies can inspire

component or cultural object classification, but the category system should not be copied directly. Local food detection can inspire industrial evidence modeling, while camera conditions, visibility, and error costs remain different. What transfers is a problem solving structure; specific assumptions must be tested again.

Visual ingredient recognition should especially not be presented as an allergen or nutritional safety guarantee. Invisible seasonings, processing, and cross contact are not fully represented in a photograph, and a dish name does not reliably establish absence. A system receiving such questions should return confirmed ingredient sources, applicable versions, and remaining unknowns, and identify who can verify the information. The later canteen blueprint addresses information organization and confirmation procedures rather than clinical or individual dietary decisions. This boundary enables the graph to perform appropriate work: reveal information gaps, locate evidence, reduce mistaken product matches, and clearly identify matters requiring human confirmation.

返回目录 Back to contents

第8章　统一证据架构与工程实现

A Unified Evidence Architecture and Engineering Design

接入层保存原始资源与版本

Preserve Original Resources and Versions at Ingestion

本章给出可实施的参考架构，属于本报告的设计综合，不声称已经在某个企业部署。接入层接收文字、页面图像、表格和业务记录，首先建立资源清单，而不是立即抽取知识。每个资源需要稳定标识、内容校验值、来源、采集时间、适用时间、权限和处理状态。同一网址内容改变时，应形成新版本；同一文件被重复上传时，应能识别重复但保留上传事件。这些基础信息决定后面能否定位证据、修复错误和删除派生数据。

解析过程应保留原始资源与标准化结果的映射。PDF 页码、图片坐标、表格单元格和视频时间戳可以采用不同定位方式，但需要一致的证据接口。解析失败时不要悄悄跳过；应记录失败类型和受影响范围，并允许部分成功。对于扫描文档，OCR 文字只是一个派生观察，原始页面仍然是核验依据。对于不断变化的业务系统，还应区分接收时间与事实生效时间，避免把晚到的数据误认为新发生的事件。接入质量是整个系统的上限之一：后续模型再强，也无法恢复已在入口丢失的页面、单位或版本条件。

This chapter proposes an implementable reference architecture as an original design synthesis, not a claim of deployment in an enterprise. Ingestion accepts text, page images, tables, and business records and first creates a resource inventory rather than immediately extracting knowledge. Each resource needs a stable identifier, content checksum, origin, collection time, applicable time, permissions, and processing status. Changed content at one URL should create a new version. Repeated uploads of one file should be recognized while retaining upload events. These foundations determine whether evidence can later be located, errors repaired, and derived data removed.

Parsing should preserve mappings between original resources and normalized outputs. PDF pages, image coordinates, table cells, and video timestamps need different locators but a consistent evidence interface. Failures should record their type and affected scope instead of silently skipping content, and partial success should be allowed. OCR text is a derived observation; the scanned page remains the verification reference. Evolving business systems also require receipt time to be distinguished from effective time so late arriving records do not appear to describe newly occurring events. Ingestion quality limits the system: later models cannot recover pages, units, or version conditions already discarded at the entrance.

观察层与实体层分开维护

Maintain Observation and Entity Layers Separately

抽取层应输出带类型的观察，而不是直接覆盖实体属性。一个观察至少包含主语候选、关系或属性、取值、限定条件、证据定位、抽取方法及审核状态。实体层负责稳定对象和身份解析，观察层负责不同来源说了什么。两层分开后，一个对象可以有相互支持或冲突的多个观察，而不必在

写入时强行选出唯一真值。面向应用的确定属性，可以由经过审核的观察生成，并保留其支持集合。

本体与抽取器之间需要版本契约。假如本体将烹饪方法拆分成预处理与加热方式，旧模型输出不能自动解释为新结构。应保存本体版本、模型版本和映射规则，并对受影响数据进行重算或显式兼容。实体消歧也不宜依赖一次不可逆批处理：先生成候选，再记录合并决定与依据，最后更新查询视图。错误被发现时，应能撤销某一次合并而不丢失原始观察。这个设计增加一些存储与管理工作，但换来的是可修复性。没有修复通道的自动构图，往往会让早期错误在后续社区摘要与回答中持续扩大。

Extraction should produce typed observations instead of directly overwriting entity properties. An observation needs at least a subject candidate, relation or attribute, value, qualifiers, evidence locator, extraction method, and review status. The entity layer manages stable objects and identity resolution; the observation layer records what different sources state. Separation permits multiple supporting or conflicting observations without forcing one truth value at write time. Confirmed application attributes can be derived from reviewed observations while retaining their support sets.

Ontology and extractor versions need an explicit contract. If an ontology divides cooking method into preparation and heating method, old outputs cannot automatically be interpreted under the new structure. Preserve ontology versions, model versions, and mapping rules, then recompute affected data or provide explicit compatibility. Entity resolution should also avoid a single irreversible batch operation. Generate candidates, record merge decisions and evidence, and update query views afterward. A discovered error should permit reversal of one merge without losing original observations. This design adds storage and management work but provides repairability. Automatic graph construction without a repair route lets early errors propagate through later community reports and answers.

图索引向量索引和资源库协同

Coordinate Graph Vector and Resource Indexes

统一架构不要求把所有数据塞进同一种数据库。资源库存放原始文件与图像，图存储表达实体、事件和证据关系，向量索引用于相似候选，关键词索引用于精确名称与编号。它们通过稳定标识连接。图数据库适合关系约束与邻域访问，向量索引适合近似相似搜索；两者解决的问题不同。部署可以从简单组合开始，只有当查询延迟、规模或运维要求实际出现时，再引入更复杂的分片和专用服务。

关键工程难点是不同索引之间的一致性。资源被撤回后，如果图边已删除而向量仍可召回旧片段，系统仍可能泄露或引用失效内容。建议以资源版本为单位维护索引状态，使用待构建、可查询、失效和重建中的明确状态，并在查询时检查证据版本。更新过程可以先生成新索引，再通过版本指针切换，避免读到半成品。对删除请求则应先阻止检索，再异步清理各派生存储，并记录清理完成证据。所谓统一，主要是身份、版本和证据契约统一，而不是存储技术统一。这个认识有助于保持架构清晰，也便于替换单个组件。

A unified architecture does not require every object to reside in one database. A resource store holds original files and images, a graph store represents entities, events, and evidence relations, a vector index

supplies similar candidates, and a keyword index supports exact names and identifiers. Stable identifiers connect them. Graph databases serve relationship constraints and neighborhood access, while vector indexes serve approximate similarity. Deployment can start with a simple combination and introduce sharding or dedicated services when measured latency, scale, or operational requirements justify them.

Consistency across indexes is the central engineering issue. If a withdrawn resource loses its graph edges but its old vectors remain retrievable, invalid content can still appear in answers. Maintain index status by resource version with explicit states for pending, queryable, invalid, and rebuilding, and check evidence versions at query time. Updates can build a new index before switching a version pointer, avoiding partially built views. Deletion should first block retrieval, then clean derived stores asynchronously while recording completion. Unification primarily concerns identity, version, and evidence contracts, rather than one storage technology. This interpretation keeps architecture understandable and allows individual components to be replaced.

查询规划形成可检查的证据包

Query Planning Produces an Inspectable Evidence Package

查询服务先识别问题中的对象、时间、范围和回答类型，再选择检索动作。明确编号可以先走精确查找，关系问题可以扩展受限邻域，主题问题可以查询社区报告，视觉问题可以调取区域或页面。路由不是不可解释的魔法分类器：系统应记录触发了什么条件，以及在不确定时采用何种回退。对于涉及多个子问题的请求，可以显式生成查询计划，逐项说明需要哪些证据，而不是让模型一次性猜出全部答案。

交给生成器的证据包应包含原文片段、来源定位、关系路径、时间条件、权限检查结果和冲突提示。原文与摘要要有不同标签，事实与推断也要区分。生成后再做陈述级核验，检查每句话是否有足够证据，引用是否指向真正支持它的内容。如果只能支持部分问题，应回答已支持的部分，并说明剩余缺口。证据包可以作为独立接口，供不同语言模型使用，也可以直接展示给人工审核员。这样系统的核心资产不只是某一次流畅回答，而是一个可复查、可比较、可重新生成的证据集合。生成模型升级时，也更容易判断变化来自表达还是来自检索。

The query service identifies objects, time, scope, and answer type before choosing retrieval actions. Exact identifiers may use direct lookup, relational questions constrained neighborhoods, thematic questions community reports, and visual questions regions or pages. Routing should record triggering conditions and fallback behavior when uncertain. A request with several subquestions can use an explicit plan stating which evidence each part requires instead of asking a model to guess the entire answer at once.

The evidence package supplied to generation should contain source passages, locators, relation paths, temporal conditions, permission checks, and conflict notices. Original material and summaries need different labels, as do facts and inferences. Claim level verification after generation checks whether each sentence has sufficient support and whether its citation points to relevant content. If evidence supports only part of a request, answer that part and explain the gap. The package can be an independent interface for different language models or for human reviewers. The durable asset becomes a reviewable, comparable, and regenerable evidence collection rather than one fluent answer. Model upgrades can then be assessed for changes in expression separately from changes in retrieval.

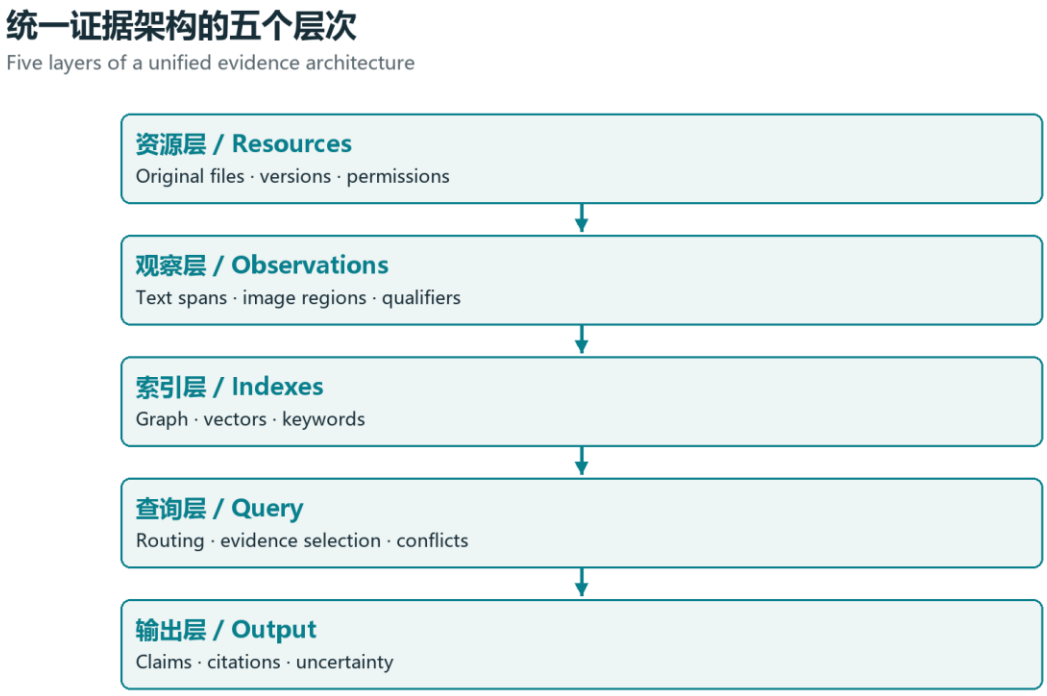


图 13 从资源到回答共享证据标识 / Shared evidence identifiers from resources to answers

来源 / Source 本报告原创设计 / Original design

增量更新摘要失效与回滚

Incremental Updates Summary Invalidation and Rollback

原始文档更新并不只影响一个片段。它可能改变实体身份、关系权重、社区归属和多个摘要，进而影响先前缓存的回答。建议把派生关系也记录成依赖图：某个摘要由哪些观察生成，某个观察来自哪些资源，某个查询缓存使用了哪些摘要。发生变化时，沿依赖关系找到需要失效的对象，而不是盲目重建全部数据或只更新最底层文件。社区重算可能产生较大范围变化，因此需要区分局部索引更新与结构性重分区。

回滚也应作为正常能力设计。模型升级后，如果抽取数量突然翻倍，可能是召回提高，也可能是把否定和假设当成事实。上线前应在固定样本上比较差异，检查高影响变化，再逐步切换。每次发布保存配置、版本、输入清单与质量摘要，使系统能够返回上一可用状态。对已经生成的答案，可以保留当时的证据快照，以解释历史输出；同时明确它是否仍代表当前知识。持续更新系统中的正确性是带时间的属性，不能靠某次离线测试永久证明。可靠运营需要让新知识进入、旧知识失效以及错误被撤销都成为可管理流程。

Updating a document can affect more than one chunk. It may change entity identity, relation weights, community membership, multiple summaries, and cached answers. Derived dependencies should therefore form a graph of their own: which observations generated a summary, which resources generated an observation, and which summaries a cached answer used. Changes can then invalidate dependent

objects rather than triggering either an indiscriminate rebuild or an insufficient file only update. Community recomputation may affect a broad region, so local index updates should be distinguished from structural repartitioning.

Rollback should be an ordinary capability. If a model upgrade doubles extracted statements, recall may have improved, or negations and hypotheses may have become facts. Compare changes on fixed samples, inspect high impact differences, and switch gradually. Preserve configuration, versions, input manifests, and quality summaries for each release so the previous usable state can be restored. Historical answers may retain evidence snapshots explaining their original output, while clearly indicating whether that evidence remains current. Correctness in an evolving system is temporal. One offline evaluation cannot prove it permanently. Operations must manage the entry of new knowledge, invalidation of old knowledge, and reversal of errors.



图 14 沿依赖链更新与撤回 / Update and withdrawal through dependencies
来源 / Source 本报告原创设计 / Original design

从原型到服务的最小交付契约

Minimum Delivery Contracts from Prototype to Service

一个可演示原型可能只需要上传文档并返回回答，一个可用服务还需要明确失败行为。输入格式不支持、权限不足、没有证据、证据冲突和模型暂时不可用，应给出不同状态，而不是统一返回空白或编造答案。接口还应提供请求标识、知识版本、引用列表和可观察的处理阶段。这样使用者能判断是稍后重试、补充材料还是提交人工核验。内部日志可以更详细，但面向用户的说明应围绕任务，不必暴露无助于决策的底层实现名词。

验收时建议使用端到端用例，覆盖一条正常查询、一条无答案查询、一条跨版本冲突、一条权限隔离和一次资源删除。每条用例都检查用户最终看到什么，而不只检查某个模块返回成功。运维指标包括错误类型、证据覆盖、处理延迟和成本，但不能用平均延迟掩盖少量极慢请求。服务契约还要规定谁负责源数据、谁审核实体合并、谁处理用户纠错。技术架构只有与这些责任关系结合，才可能持续运行。本报告的架构图因此同时展示数据流和反馈回路：系统不仅从文档生成答案，也从被发现的问题返回源数据与知识维护流程。

A demonstration may only upload documents and return answers. A usable service must define failure behavior. Unsupported formats, insufficient permissions, absent evidence, conflicting evidence, and temporary model unavailability should produce distinct states instead of one blank response or an invented answer. Interfaces should include a request identifier, knowledge version, citation list, and observable processing stages. Users can then decide whether to retry, add material, or seek human verification. Internal logs may be detailed, while user explanations should focus on the task rather than implementation terms that do not help decisions.

Acceptance testing should cover an ordinary query, an unanswerable query, a version conflict, permission isolation, and resource deletion. Each case examines what the user ultimately sees rather than only whether one module succeeds. Operational measures include error types, evidence coverage, latency, and cost, with attention to slow requests hidden by averages. Service contracts also define who owns source data, reviews entity merges, and handles corrections. Architecture becomes sustainable only when connected to these responsibilities. The reference diagram therefore includes both data flow and feedback: the system transforms documents into answers and returns discovered problems to source and knowledge maintenance processes.

返回目录 Back to contents

第9章　分层评测实验设计与不确定性

Layered Evaluation Experimental Design and Uncertainty

先定义任务与数据划分

Define Tasks and Data Splits First

评测首先回答系统要帮助谁完成什么工作。事实查询、跨文档关系发现、主题概括和视觉定位需要不同的成功标准，不能只用一个通用问答分数覆盖。建议为每个问题记录用户意图、可接受答案、必要证据、不可接受推断和无答案条件。标注者应能看到这些规则，并在存在争议时记录理由。对于开放式概括，可以定义需要覆盖的主题和必须保留的分歧，不必假设只有一段唯一正确文本。评测综述和 RAGAs 提供不同层次的评估思路，但具体量表仍需适配本地任务。[R25][R26]

数据划分应阻断与目标部署不一致的信息泄漏。若系统将服务新门店，可按门店划分；若服务未来资料，可按时间划分；若要检验新实体泛化，可把相同实体的近重复资料放在同一侧。训练、提示开发和最终测试也需要区分。测试问题如果由同一批社区摘要自动生成，再用这些摘要作检索材料，可能过度奖励摘要自身的措辞。应加入独立编写的问题和用户真实问题，并把问题来源作为分析维度。评测集不是装饰性附件，它明确了结论可以支持的使用范围；范围之外的迁移需要新测试。

Evaluation begins with who needs help and what work they need to complete. Factual lookup, relationships across documents, thematic synthesis, and visual localization require different success criteria. One generic question answering score is insufficient. Each question should specify intent, acceptable answers, necessary evidence, forbidden inferences, and conditions for no answer. Annotators

need these rules and should record reasons for disagreements. Open synthesis can be evaluated through required themes and preserved disagreements without assuming one uniquely correct paragraph. Evaluation surveys and RAGAs offer useful approaches, while local tasks still require adapted rubrics.[R25][R26]

Splits should prevent leakage inconsistent with deployment. New restaurant use can require restaurant based splits, future use temporal splits, and new entity generalization entity separated near duplicates. Training, prompt development, and final testing should be distinct. Questions generated from community summaries and evaluated using those same summaries may overreward their wording. Independently authored and real user questions should be included, with question origin retained as an analysis variable. The evaluation set defines the range of use supported by a conclusion. Transfer beyond that range requires further testing.

检索指标要匹配相关性的定义

Match Retrieval Metrics to Relevance Definitions

精确率衡量返回结果中有多少相关，召回率衡量全部相关证据中找回多少。二者都需要先定义相关证据集合；当标注不完整时，未标注结果不能直接视为无关。对于分级相关性，可使用 nDCG 衡量较重要证据是否排在前面。其计算依赖增益函数、折扣与截断位置，报告时必须写清。本文演示采用增益为二的相关等级次幂减一、按排名加一的二进对数折扣，理想排序来自全部已标注候选，不能只重排已返回项。不同定义的 nDCG 不能直接混在同一张比较表里。

图检索还应检查路径层次的完整性。找到了两个端点却没有连接证据，不等于完成多跳问题；找到了路径但遗漏关键的否定条件，也不能算完整支持。可为每个问题标注最小证据集合或若干可替代集合，并统计是否至少覆盖其中一组。全局主题问题则需要覆盖不同来源和观点，避免高频重复片段占满前若干结果。建议同时记录重复率、来源多样性和无关扩展比例。指标越接近实际证据需求，越能解释为什么一套看似召回很多的系统仍然回答不好。排名质量是中间目标，最终还要看证据能否支持任务。

Precision measures relevant items among returned results, while recall measures recovered items among all relevant evidence. Both require a defined relevance set. With incomplete annotation, an unjudged result should not automatically count as irrelevant. Graded relevance can use nDCG to assess whether more important evidence appears earlier. Its gain function, discount, and cutoff must be stated. The demonstration here uses exponential gain, two raised to the relevance grade minus one, and a base two logarithmic discount on rank plus one. The ideal ranking uses all judged candidates, not only a reordering of retrieved results. Scores with different definitions should not be combined in one comparison table.

Graph retrieval also requires path completeness. Finding both endpoints without connecting evidence does not solve a multihop question. Finding a path while omitting a critical negation does not provide complete support. Questions can have a minimum evidence set or several interchangeable sets, with success requiring coverage of at least one. Global themes additionally require source and viewpoint coverage so repeated high frequency passages do not occupy every top position. Duplicate rate, source diversity, and irrelevant expansion should accompany ranking measures. Metrics aligned with evidence needs explain why apparently extensive retrieval can still produce poor answers. Ranking is an intermediate objective; evidence must ultimately support the task.

$$Precision@k = \frac{\text{Relevant retrieved within k}}{k}$$

$$Recall@k = \frac{\text{Relevant retrieved within k}}{\text{All relevant evidence}}$$

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

$$nDCG@k = \frac{DCG@k}{IDCG@k}$$

公式中的相关集合与理想排序必须固定 / Fix relevance sets and ideal rankings.

回答质量与引用质量分开判断

Assess Answer Quality and Citation Quality Separately

一段回答可能写得正确却引用错误，也可能忠实复述一份本身有误的材料。因此至少需要区分事实正确性、对给定证据的忠实性、问题覆盖程度和引用支持。可将答案拆成可核验陈述，为每条陈述标记直接支持、部分支持、矛盾或证据不足。引用精确率关注被引用材料是否真正支持相应陈述，引用覆盖关注需要证据的陈述是否得到支持。只检查引用链接能否打开，无法说明内容支持关系成立。对于摘要产生的答案，应继续追踪到原文，而不止停在另一段模型摘要。

评审还要保留不确定性的质量。问题证据不足时，明确承认未知往往优于完整但无依据的答案；证据冲突时，列明不同来源及适用时间可能优于强行综合。可以将合理拒答与过度拒答分别计数，避免系统靠拒绝所有问题获得高安全分。中英对照报告或跨语言问答还应检查数字、否定、量词和条件是否一致，因为翻译流畅并不保证事实边界相同。一个有用的审核样本应同时包含容易回答、部分可答、存在冲突和完全无答案四类问题。只有覆盖这些情形，才能评估系统是否知道何时应该停止推断。

An answer can be correct while citing the wrong material, or faithfully repeat a source that is itself mistaken. Evaluation should distinguish factual correctness, faithfulness to supplied evidence, question coverage, and citation support. Break answers into verifiable claims and classify each as directly supported, partially supported, contradicted, or unsupported. Citation precision asks whether cited material supports the associated claim. Citation coverage asks whether claims requiring evidence receive support. A working link alone does not establish support. Answers generated through summaries should remain traceable to original material rather than ending at another model summary.

Uncertainty handling also needs assessment. Admitting missing evidence may be better than giving a complete unsupported response. Conflicting sources may require separate statements with temporal scope rather than forced synthesis. Count justified abstention and excessive abstention separately so refusing everything does not earn a misleading quality score. Bilingual reports and cross language answers should compare numbers, negations, quantifiers, and conditions because fluent translation can change factual boundaries. An effective review sample includes answerable, partially answerable, conflicting, and unanswerable questions. These conditions reveal whether a system recognizes when inference should stop.



图 15 单一总分不能定位失败来源 / One score cannot locate failures

来源 / Source 本报告原创设计 / Original design

多模态贡献必须用对照证明

Demonstrate Multimodal Contribution through Controls

多模态评测需要确认问题真的要求多种模态。有些标称视觉问答的数据，仅凭标题或文字就能猜中答案；这种情况下，即使多模态系统得分高，也不能证明它读懂了图像。MultiModalQA 与 WebQA 提供跨模态复杂问答的研究实例，Visual-RAG 进一步关注视觉知识密集查询。采用这些基准时，仍应检查具体版本、任务和证据标注，不把数据集名称当成视觉必要性的充分保证。[R30][R31][R21]

建议设计文字单独、图像单独、完整融合和错误配对四种条件，并保持生成模型与预算可比。如果完整融合优于单独模态，同时错误配对明显损害相关问题，才更有理由认为系统使用了跨模态联系。还可以遮蔽答案区域、移除表头或替换相似图片，检查回答是否相应改变。对图结构，可比较无图、随机边、真实边和只保留实体名四种条件。所有消融都应说明哪些组件实际运行，哪些只是计划。本报告不报告未执行的模型消融分数，而是提供可用于后续研究的实验模板。缺少执行结果时，明确假设比绘制虚构性能柱状图更有研究价值。

Multimodal evaluation must establish that questions actually require multiple modalities. Some nominally visual questions can be guessed from titles or text alone. High scores then do not prove image understanding. MultiModalQA and WebQA offer research examples of complex multimodal questions, while Visual-RAG examines visually knowledge intensive queries. Using these benchmarks still requires checking versions, tasks, and evidence annotations; a dataset name does not guarantee visual necessity.[R30][R31][R21]

Useful conditions include text alone, image alone, complete fusion, and incorrect pairings, with comparable generators and budgets. If fusion improves on individual modalities and incorrect pairings damage the relevant questions, there is stronger evidence of cross modal use. Masking answer regions, removing headers, and substituting similar images can test whether answers change appropriately. Graph controls can compare no graph, random edges, real edges, and entity names alone. Every ablation should distinguish executed components from planned ones. This report supplies experimental templates rather than scores for model ablations that were not run. Explicit hypotheses without results are more useful than invented performance bars.

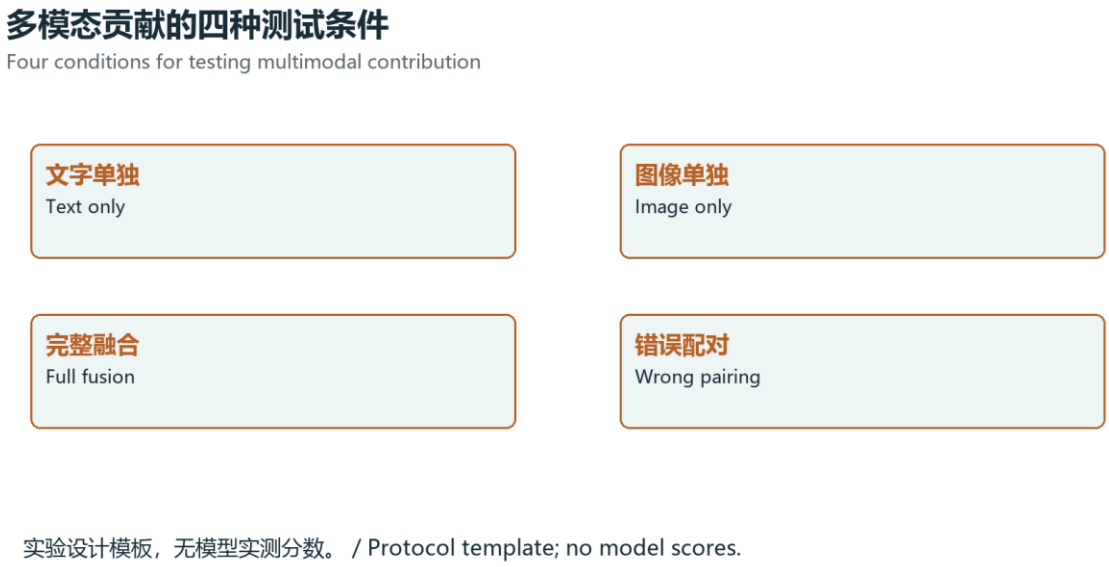


图 16 用对照检查视觉信息是否真正被使用 / Controls for actual visual contribution
来源 / Source 本报告原创设计 / Original design

自动评审需要校准与人工复核

Calibrate Automated Judges and Retain Human Review

语言模型评审能降低大规模人工检查的负担，但它也是测量工具，需要验证。评审器可能偏好更长、更流畅或与自身表述相近的答案，也可能被错误引用的外观说服。应先由人工对一组样本建立判断，再比较自动评审在不同错误类型上的一致性，而不只看总体相关性。对低一致性类别，可以改写量表、增加证据定位或直接保留人工判断。RAGAs 的自动化思路适合构建反馈回路，但自动分数不能独自承担最终正确性的证明。[R25]

评审提示应明确允许不知道，要求引用支持位置，并避免把系统名称或作者身份暴露给评审器。比较两套答案时，可随机交换顺序并检查顺序效应。人工标注也不是绝对真值：需要多人独立标注、分歧讨论和规则修订，特别是开放摘要中的重要性判断。对高影响错误，应保存案例而不只保留分数，让团队能看到错误如何产生。报告结果时可以同时提供自动指标、人工抽查比例与主要不一致类型。这样读者能判断测量的可靠程度，并理解为什么某个分数变动还不足以支持上线或替换现有系统。

Language model judges can reduce large scale review effort, but they are measurement instruments requiring validation. A judge may prefer longer, more fluent answers or wording resembling its own, and may be persuaded by citations that only look convincing. Human judgments on a sample should first establish a reference, followed by agreement analysis across error types rather than only overall correlation. Low agreement categories may need revised rubrics, better evidence locations, or continued human assessment. RAGAs supports automated feedback, but an automated score alone cannot prove final correctness.[R25]

Judge instructions should permit uncertainty, require supporting locations, and conceal system names or author identities. Pairwise comparisons can randomize answer order and test order effects. Human annotation is not absolute truth either. Independent judgments, disagreement discussions, and rule revision are especially important for importance judgments in open summaries. High impact errors should retain cases rather than only scores so teams can inspect how failures occur. Results can report automated measures, human sampling rates, and major disagreement types together. Readers can then assess measurement reliability and understand why one score change may not justify deployment or replacement.

报告差异分布而不只报告平均值

Report Distributions and Differences beyond Averages

实验结果应按问题类型、来源、实体频率和证据长度分组。平均提升可能来自少数容易问题，也可能掩盖某类重要问题的退步。对于同一测试问题上的两套系统，宜分析成对差异，并报告样本量和不确定性。若使用重采样估计区间，应说明重采样单位；当多个问题来自同一文档时，把每题视为完全独立可能低估不确定性。多次尝试后只报告最好一次，也会使结果显得过于稳定。模型随机性、索引随机性和标注差异应在研究设计中分别考虑。

统计显著与实际价值不是同一判断。一个很小但稳定的分数提升，可能不足以承担额外构图成本；一个总体不显著的变化，也可能对少数高价值关系问题很重要。建议把效果大小、成本、错误严重性和维护负担放在同一决策表中。数据不足时，应报告证据仍不充分，并设计下一轮更有辨别力的实验，而不是用复杂统计掩饰薄弱样本。本报告的演示排名只用于说明计算方法，不能估计真实系统效能；业务案例中的百分比也不在缺少原始样本时被赋予虚构区间。清楚报告测量边界，是让研究能够被继续验证的必要条件。

Results should be grouped by question type, source, entity frequency, and evidence length. Average gains can arise from easy questions while hiding regressions on important categories. Systems evaluated on the same questions should use paired differences and report sample size and uncertainty. Resampling intervals require an explicit sampling unit. Treating every question as independent when several share a document can underestimate uncertainty. Reporting only the best of many attempts also exaggerates stability. Model randomness, indexing randomness, and annotation differences should be considered separately.

Statistical significance and practical value are different judgments. A small stable score gain may not justify additional construction cost, while an uncertain average change may matter for a small set of valuable relational tasks. A decision table should combine effect size, cost, error severity, and maintenance burden. Insufficient data should lead to an explicit uncertainty statement and a more discriminating next experiment rather than elaborate statistics over weak samples. The demonstration

ranking in this report explains calculation only and does not estimate system effectiveness. Business percentages also receive no invented intervals without original samples. Reporting these boundaries enables further verification.

返回目录 Back to contents

第10章　成本性能与持续运行

Cost Performance and Continuous Operation

将构建查询和维护成本分别计算

Separate Construction Query and Maintenance Costs

图增强系统的成本不能只看一次回答消耗多少模型调用。构建阶段包含解析、抽取、实体消歧、图处理和摘要；查询阶段包含路由、检索、重排、生成与核验；维护阶段还包含更新、纠错、权限变更和人工审核。本报告建议以固定观察周期记录三部分，并明确哪些是一次性支出、哪些随文档增加、哪些随请求增加。只有这样才能判断某种较昂贵索引是否会在大量查询中获得合理摊销。

一个简单的成本模型可以写成周期总成本等于构建成本、更新成本、查询次数乘平均查询成本以及运维成本之和。模型中的每个量都要有单位和测量边界：模型用量、计算时间、存储和人工工时不能未经换算直接相加。若使用货币估算，应记录价格日期与计价来源；本报告不引用易变的外部模型报价，而在演示中使用明确标注的假设成本单位。平均每次查询成本还会随问题类型变化，主题概括可能需要更多报告，精确查询可能只需一个片段。因此成本分析应与任务分布相连，避免用一种复杂查询代表全部请求。

The cost of a graph augmented system extends beyond model calls for one answer. Construction includes parsing, extraction, entity resolution, graph processing, and summaries. Querying includes routing, retrieval, reranking, generation, and verification. Maintenance includes updates, corrections, permission changes, and human review. Record all three over a defined period, distinguishing one time expenditure from costs growing with documents or requests. This makes it possible to judge whether an expensive index can be reasonably amortized across many queries.

A simple period cost equals construction, updates, query count multiplied by average query cost, and operations. Every quantity needs units and measurement boundaries. Model usage, compute time, storage, and human hours cannot be added without conversion. Monetary estimates require dated prices and their sources. This report instead uses explicitly assumed cost units for demonstrations, avoiding unstable external quotations. Query cost also varies by task. Thematic synthesis may require many reports while precise lookup may require one passage. Cost analysis should follow the task distribution rather than treating a complex query as representative of all requests.

$$C_{\text{total}} = C_{\text{build}} + C_{\text{update}} + N_{\text{query}} \times C_{\text{query}} + C_{\text{operations}}$$

延迟预算沿查询路径分配

Allocate Latency Budgets across the Query Path

用户感受到的是端到端等待时间，而不是某个检索模块的单独耗时。查询路径中，权限检查、网络传输、图扩展、图片读取和生成都可能成为瓶颈。建议先测量各阶段分布，再决定优化位置。若大部分时间用于生成，优化毫秒级图查询未必改变体验；若少量大图查询导致长尾，平均延迟也会掩盖问题。报告中应同时记录中位数和较高分位，并按任务类型区分。超时后返回部分结果还是明确失败，也应作为服务设计的一部分。

预算可以在执行前分配，也可以随证据发现动态调整。例如，精确编号已经找到一份当前有效记录，就不一定继续扩展整个关系网络；关系问题缺少一个关键连接时，则可把更多预算分配给定向查找。终止规则应依据证据充分性，而不只是固定步数。为防止图扩展无限增长，可以设置候选上限、关系类型白名单和时间范围。预算不足时，系统应说明当前证据覆盖到哪里，而不是把未搜索区域当成不存在。性能优化因此涉及回答范围与可靠性的共同选择，不能只追求更快输出文字。

Users experience end to end waiting time rather than one retriever's isolated duration. Permission checks, network transfer, graph expansion, image loading, and generation can all become bottlenecks. Measure stage distributions before selecting optimization targets. If generation dominates, improving a millisecond graph query may not change the experience. If rare large traversals create a long tail, average latency hides the issue. Record medians and higher percentiles by task type, and define whether timeouts produce partial results or explicit failure.

Budgets can be assigned before execution or adjusted as evidence appears. An exact identifier yielding one current authoritative record may not require further network expansion. A relational question missing one crucial connection may justify more directed retrieval. Stopping rules should reflect evidence sufficiency rather than only a fixed number of steps. Candidate limits, allowed relation types, and temporal scope prevent unbounded expansion. When a budget is exhausted, explain the evidence coverage achieved instead of treating unsearched regions as absent. Performance optimization therefore selects answer scope and reliability together, rather than merely accelerating text production.

缓存必须绑定知识和权限版本

Bind Caches to Knowledge and Permission Versions

缓存可以减少重复处理，但也可能重复传播过期或越权信息。查询字符串相同并不意味着可复用同一答案，因为用户权限、知识版本和时间条件可能不同。建议缓存键包含规范化问题、资源范围、知识版本、模型配置和权限范围标识。缓存内容还应保存其依赖的证据集合，以便源数据改变时失效。对于全局社区摘要，尤其需要检查它是否混合了不同访问级别的信息；仅在最终回答阶段过滤原文，无法消除摘要中已泄露的内容。

缓存可以分层设置：解析结果按资源内容复用，嵌入按模型与内容版本复用，检索候选按范围复用，最终答案则采用更严格条件。不同层的失效规则不应混成一个统一有效期。资源内容不变但权限收紧时，解析缓存可能继续有效，答案缓存却必须重新检查；模型升级时，旧嵌入可能需要

重算，原始资源不必复制。测试应覆盖这些交叉变化。缓存命中率高并不自动意味着系统更好，如果命中的是过时或不完整回答，反而会稳定地产生错误。因此应同时观察缓存正确性、节约成本和失效延迟，而不是只把命中率作为成功指标。

Caching reduces repeated work but can also repeatedly distribute outdated or unauthorized information. Identical query text does not guarantee answer reuse because permissions, knowledge versions, and temporal conditions may differ. Cache keys should include normalized questions, resource scope, knowledge version, model configuration, and permission scope. Cached objects should retain evidence dependencies for invalidation when sources change. Global summaries require special care if they combine access levels. Filtering original passages at the final answer stage cannot remove information already exposed inside a summary.

Different layers can cache parsing by content, embeddings by model and content version, retrieval candidates by scope, and answers under stricter conditions. Their invalidation rules should not collapse into one expiration period. Tightened permissions may leave parsing reusable while requiring answer revalidation. A model upgrade may require new embeddings without duplicating original resources. Tests should cover these combinations. High hit rates do not automatically indicate a better system: cached stale or incomplete answers can produce errors consistently. Cache correctness, cost savings, and invalidation delay should accompany hit rate as measures.

压缩和剪枝需要保留关键例外

Preserve Critical Exceptions during Compression and Pruning

为了减少上下文和存储，可以压缩片段、合并重复观察、剪去低价值边或选择较粗社区层级。风险在于重要性不能只由频率或长度决定。一句很短的例外条件可能改变整段规则，一条低频关系可能是某次事故的唯一证据。压缩前应识别否定、数量、单位、时间和例外，并检查压缩后是否仍能恢复这些约束。对于无法安全概括的内容，可以保留原文并降低其他背景信息的优先级。压缩质量应通过任务结果与证据保留共同评估。

图剪枝也应考虑证据角色。外围节点虽然不参与最短路径，却可能提供反例或解释身份差异；删除它们可能让图看起来更清晰，却让回答变得单向。可区分用于导航的精简图与用于核验的证据图，前者帮助快速定位，后者在需要时补充完整出处。两者通过标识关联，不必把所有信息都放进模型上下文。对大量重复来源，应合并展示但保留来源计数与独立性判断，避免把转载次数当作多个独立证据。有效压缩减少冗余，保留能改变结论的差异，这比单纯追求更少节点或更短提示词更接近实际目标。

Reducing context and storage can involve compressing chunks, merging repeated observations, pruning edges, or selecting coarser communities. Frequency and length are poor substitutes for importance. A short exception can reverse a rule, and a rare relation may be the only evidence for an incident. Compression should identify negation, quantities, units, time, and exceptions, then verify their preservation. Material that cannot be safely summarized can retain original wording while less important background receives lower priority. Compression quality requires both task outcomes and constraint preservation.

Graph pruning also needs evidence roles. Peripheral nodes absent from a shortest path may contain counterexamples or identity distinctions. Removing them can make a graph visually cleaner but answers more one sided. A compact navigation graph can support fast discovery while a linked evidence graph retains verification material. Stable identifiers connect them without placing every object in model context. Repeated sources can be combined in presentation while preserving counts and independence assessments, so republication does not masquerade as independent support. Useful compression removes redundancy while retaining differences capable of changing a conclusion, rather than merely minimizing nodes or prompt length.

观测质量漂移与运营反馈

Observe Quality Drift and Operational Feedback

上线后，数据和用户问题都会变化。新术语、新拍摄设备或新文档格式可能使抽取和检索质量下降，而模型服务本身仍返回成功。应建立定期抽样与固定哨兵问题，监测实体未链接率、无证据回答、引用失效和模态缺失等信号。指标变化需要结合输入分布解释，不能把所有下降归咎于模型。若用户开始提出更复杂的问题，平均成功率变化可能反映任务变了。监控应能区分数据漂移、任务漂移和系统故障。

反馈入口要允许用户指出具体陈述或证据的问题，而不只是给整个回答点赞。纠错记录应回到知识维护队列，标记是源数据错误、抽取错误、身份合并错误、检索遗漏还是生成越界。高频错误可以形成回归用例，但需要避免把全部用户反馈直接当作正确标签。对于争议性反馈，应保留原始意见与审核结果。运营团队还应观察纠错处理时间和同类错误复发率，因为一个能发现问题却长期不修复的系统仍然不可靠。持续改进的闭环，是从用户看到的错误回到可修改组件，再通过独立检查确认修复生效。

Data and user questions change after deployment. New terminology, cameras, and document formats can degrade extraction or retrieval while model services continue returning successful responses. Periodic sampling and fixed sentinel questions can monitor unresolved entities, unsupported answers, invalid citations, and missing modalities. Changes need interpretation alongside input distributions rather than automatic attribution to models. More complex user questions can lower average success because the task changed. Monitoring should distinguish data drift, task drift, and system failures.

Feedback should let users identify a problematic claim or source rather than only rate an entire answer. Corrections enter a maintenance queue classified by source errors, extraction errors, identity merges, retrieval omissions, or generation overreach. Repeated failures can become regression cases, while user feedback should not automatically become a correct label. Disputed feedback needs both the original comment and review outcome. Teams should measure correction time and recurrence, since detecting problems without repairing them does not make a system reliable. The improvement loop connects visible errors to editable components and confirms repairs through independent checks.



图 17 用户反馈到组件修复的闭环 / Feedback to component repair
来源 / Source 本报告原创设计 / Original design

采用门槛应与价值和维护能力匹配

Match Adoption Thresholds to Value and Maintenance Capacity

采用 GraphRAG 不必从全量资料和全部功能开始。可以选择一个边界清楚、问题重复出现、证据可获得的场景，先验证图结构是否减少关键错误或提高查找效率。成功标准应在实验前写明，例如关系问题的必要证据覆盖、审核时间和错误严重性。若只在演示中挑选系统擅长的问题，就无法判断真实价值。试点还需要包含数据更新和纠错过程，因为静态演示最容易遗漏维护负担。能够持续修复一张小图，比一次性生成一张巨大但不可管理的图更接近可用服务。

决策可以分阶段进行：证明数据可组织，证明检索有增量，证明用户任务受益，最后证明持续运营可承担。每一阶段都允许停止或简化。如果向量混合检索已经满足目标，图可以只保留用于解释或特定关系查询；如果视觉信息只对少量问题必要，则可按需调用。团队能力也属于成本的一部分，包括本体维护、数据审核和故障处理。报告建议把退出机制写进试点方案：如何导出证据、替换模型和回到前一流程。可替换性减少长期依赖，使技术选择能够随新证据调整，而不是被早期架构承诺锁定。

GraphRAG adoption need not begin with every document and feature. Choose a bounded scenario with recurring questions and available evidence, then test whether graph structure reduces important errors or improves search efficiency. Define success beforehand through necessary evidence coverage, review time, and error severity. Selecting only questions a demonstration handles well cannot establish real value. Pilots should include updating and correction because static demonstrations conceal maintenance burdens. A small graph that can be repaired continuously is closer to a usable service than a huge unmanageable graph generated once.

Adoption can proceed through evidence that data can be organized, retrieval gains are real, user tasks benefit, and operations are affordable. Each stage permits stopping or simplifying. If hybrid vector retrieval meets the goal, graphs may remain only for explanations or selected relational queries. Visual processing can be invoked only where necessary. Team capabilities in ontology maintenance, data review, and incident handling are part of cost. Pilots should define exit routes for exporting evidence, changing

models, and restoring earlier workflows. Replaceability allows decisions to follow new evidence rather than becoming locked into early architectural commitments.

返回目录 Back to contents

第11章　权限治理安全与责任

Permissions Governance Security and Responsibility

来源权威与知识状态分别表达

Separate Source Authority from Knowledge Status

图谱中的知识不应只分为存在与不存在。可区分原始记录、模型观察、审核确认、存在争议和已撤回等状态，并保存状态变化的原因。来源权威也不是一个适用于所有问题的单值：制造商手册可能适合设备参数，维修记录更适合说明实际更换，用户评论则反映使用经历。评估来源时，应问它对哪一类陈述有资格提供证据，以及是否覆盖当前对象与时间。来源声誉不能替代陈述级核验，多个低质量转载也不等于多个独立确认。

溯源模型可以借鉴 W3C PROV 中实体、活动和责任主体之间的关系，用来表达数据如何被处理以及由谁产生。采用这种思路不意味着必须将整个标准一次性实现；最小系统也可以先记录资源、处理步骤和责任人，再逐步增加细节。[R28]关键是让派生知识具有可解释的历史，而不是只存最终属性。面对冲突时，系统应能展示不同来源及其适用范围，必要时请求人工裁决。治理的目标不是把不确定信息全部排除，而是让不同知识状态对检索、生成和用户判断产生明确影响。

Knowledge in a graph should have more states than present or absent. Original records, model observations, reviewed confirmations, disputed claims, and withdrawn claims can retain reasons for state changes. Source authority is also question dependent. A manufacturer manual may support specifications, maintenance records actual replacements, and reviews user experiences. Ask which claims a source is qualified to support and whether it covers the current object and period. Reputation does not replace claim level verification, and several poor republications do not establish independent confirmation.

Provenance design can draw on W3C PROV relations among entities, activities, and responsible agents to describe how data was processed and who produced it. This does not require implementing the entire standard at once. A minimal system can first record resources, processing steps, and responsibility, then extend detail.[R28] The essential property is an explainable history rather than only a final attribute. Conflicts should expose sources and scope and, when necessary, request human judgment. Governance should allow uncertain information to remain useful while ensuring its status affects retrieval, generation, and interpretation explicitly.

权限必须沿派生链传播

Propagate Permissions through Derived Information

图谱可能把原本分散的资料连接起来，产生比单份文档更敏感的信息。即使每条边本身不显眼，组合后也可能暴露人员关系、业务计划或未公开项目。因此权限检查不能只在下载文件时执行，而应覆盖节点、观察、路径、摘要和最终答案。最安全的设计起点是：检索只能看到当前用户有权使用的证据，派生对象的可见范围由其依赖来源共同决定。具体策略需要结合组织制度，不能靠一个统一公开标记解决。

社区摘要是容易忽视的环节。如果一个摘要混入受限来源，向无权限用户展示其概括仍可能泄露内容，哪怕原始引用被隐藏。可以按权限范围分区生成摘要，或在查询时使用经过权限过滤的证据重新概括，但两种方案各有成本。测试应采用成对用户：同一问题在不同权限下返回不同可用证据，同时都不能暗示隐藏资料的具体内容。权限变更后还要检查缓存与历史共享链接。治理评估不仅检查直接泄露，也检查关系组合产生的推断风险，并由了解业务的人员判断哪些关联需要限制展示。

A graph can connect scattered records into information more sensitive than any individual document. Unremarkable edges may jointly reveal relationships, plans, or undisclosed projects. Permission checks should therefore cover nodes, observations, paths, summaries, and answers, not only file downloads. A sound starting rule is that retrieval sees only evidence the current user may use, while the visibility of derived objects depends on their source dependencies. Actual policy requires organizational decisions rather than one public flag.

Community reports are easy to overlook. A summary containing restricted sources can leak their substance even if original citations are hidden. Possible strategies include generating summaries within permission partitions or synthesizing filtered evidence at query time, each with different costs. Tests should use paired users asking the same question under different permissions. Both should receive only allowed evidence without revealing the substance of hidden records. Permission changes also require cache and historical link checks. Governance should examine direct leakage and risks arising from combined relations, with domain staff deciding which associations require restricted display.

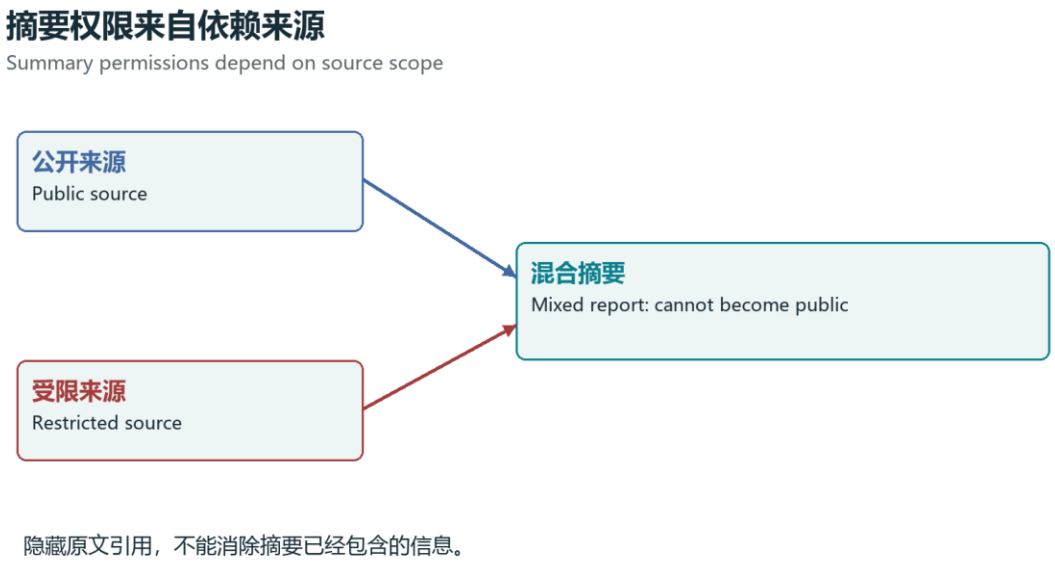


图 18 权限应覆盖派生摘要而不只覆盖原文 / Permissions must cover derived reports
来源 / Source 本报告原创设计 / Original design

检索内容是资料而不是执行指令

Retrieved Content Is Data rather than Execution Authority

文档、网页和图片中的文字可能包含命令式内容，例如要求忽略规则、透露其他资料或执行外部操作。即使这些文字被检索到，也只是待分析材料，不会自动获得系统指令权限。多模态系统还可能从图片中的文字或隐藏版面中读到类似内容。应在架构中区分用户请求、应用规则和外部资料，并让生成器知道引用材料只能作为证据。抽取器也不能因为原文要求“新增某关系”就把它当作事实，而应判断该文本是在描述、引用还是发出指令。

防护需要覆盖工具调用边界。检索到的文档不能直接授权发邮件、修改数据库或下载执行文件；这些动作应由用户任务和应用权限决定。可把生成与执行分开，先形成可检查计划，再由受限接口处理已授权动作。测试材料应包含显式诱导、伪装成系统提示的段落以及图像内文字，并检查系统是否仍按原任务回答。日志中记录异常内容的位置有助于调查，但不要把敏感材料无控制地复制到更多日志。这个原则同样适用于研究写作：来源中的表达可以被总结和批判，不能代替用户对研究范围与交付的要求。

Documents, webpages, and images may contain commands to ignore rules, reveal other information, or perform external actions. Retrieval does not grant those commands authority; they remain material to analyze. Multimodal systems can encounter similar instructions inside images or unusual layouts. Architecture should distinguish user requests, application rules, and external material, and treat quoted content as evidence. Extractors should also distinguish description, quotation, and instruction instead of creating a relation merely because a source tells the system to add one.

Protection must extend to tool boundaries. Retrieved documents cannot authorize email, database changes, or downloaded execution. User tasks and application permissions determine such actions. Generation and

execution can be separated through inspectable plans and constrained interfaces for authorized operations. Tests should include explicit inducements, paragraphs pretending to be system messages, and instructions embedded in images, then check continued adherence to the original task. Logging suspicious locations helps investigation, but sensitive material should not spread into uncontrolled logs. The same principle governs research writing: sources may be summarized and criticized, but they do not replace the user's research scope or delivery requirements.

删除更正与可追踪的生命周期

Deletion Correction and a Traceable Lifecycle

知识图谱的删除不是移除一个文件名。原始资源可能已经产生文字片段、向量、实体观察、关系、社区报告和答案缓存。若不记录依赖，系统很难知道还剩哪些副本。建议以资源和观察标识建立派生清单，删除时先撤销可查询状态，再逐层清理，并形成完成记录。对于需要保留审计信息的场景，应由组织明确哪些最小元数据可以保留，避免技术团队自行把全部原始内容复制进所谓审计备份。这里讨论工程流程，不替代具体地区或机构的规则判断。

更正与删除也应区分。某条观察的取值被修正时，可能保留旧版本以解释历史答案，但当前查询必须使用有效版本；某份资源被撤回时，则不能继续让其派生摘要支撑新回答。用户界面可以显示更正日期与影响范围，让读者知道旧引用为何失效。测试时应在删除前后执行相同查询，并检查原文、图路径、向量召回和缓存答案是否都符合预期。若外部系统保留了副本，还要说明系统能够控制的边界。可靠生命周期管理的标准，是能证明信息如何进入和退出，而不是只声明支持删除。

Deleting from a graph system means more than removing a filename. A resource may have produced chunks, vectors, entity observations, relations, community reports, and answer caches. Without dependencies, remaining copies are hard to identify. Resource and observation identifiers should support a derived object inventory. Deletion first revokes queryability, then cleans dependencies and records completion. Where auditing requires retention, organizations should define the minimum metadata allowed rather than letting teams copy all content into an audit backup. This is an engineering process, not a substitute for jurisdictional or institutional policy decisions.

Correction differs from deletion. A corrected observation may retain an earlier version to explain historical answers while current queries use the effective version. A withdrawn resource should no longer support new answers through its summaries. Interfaces can show correction dates and affected scope so readers understand invalid citations. Tests should repeat queries before and after deletion and inspect original content, graph paths, vector retrieval, and cached answers. External copies require an explicit statement of control boundaries. Reliable lifecycle management demonstrates how information enters and leaves, rather than merely advertising a delete feature.

人工审核关注高影响节点

Focus Human Review on High Impact Decisions

所有知识都人工审核通常不可承担，完全自动处理又容易积累错误。可以按影响分配审核资源：被大量答案使用的观察、连接多个业务对象的实体合并、与现有记录冲突的更新以及决定用户行动的属性，应优先检查。影响不仅取决于图中心性，还取决于错误后果与使用频率。一条连接度很低的安全相关陈述，也可能比热门主题标签更值得审核。审核队列需要解释为什么这条记录被选中，以及审核者需要核对哪些证据。

审核界面应并排展示原文、图片区域、当前候选和历史变化，允许确认、拒绝、拆分实体或要求补充材料。不要只显示模型给出的置信分数，因为分数可能没有校准，也可能来自不同模型而不可比较。人工决定应成为新的审计记录，并与模型输出区分。对于重复出现的同类错误，可以修订规则或数据流程，而不是无限增加人工负担。审核质量也需要抽查，防止审核者因自动化建议而机械确认。有效的人机协作不是让人给每条输出盖章，而是让人在能够改变结果的关键节点上获得足够信息和实际修改权。

Reviewing all knowledge manually is often unaffordable, while complete automation accumulates errors. Review effort can follow impact: observations supporting many answers, entity merges connecting business objects, conflicting updates, and action determining attributes deserve priority. Impact depends on consequences and usage, not only graph centrality. A low degree safety related statement may merit more attention than a popular topic label. Review queues should explain selection reasons and the evidence requiring inspection.

Review interfaces should show original text, image regions, candidates, and change history together, with actions to confirm, reject, split entities, or request more material. Model confidence alone is insufficient because scores may be uncalibrated or incomparable across models. Human decisions should create separate audit records. Repeated error types can motivate rule or process changes instead of indefinitely increasing manual work. Review quality itself needs sampling to detect mechanical acceptance of automated suggestions. Effective collaboration gives people adequate information and real correction power at decisions where they can change outcomes, rather than asking them to stamp every output.

事故处理与责任归属可操作化

Make Incident Response and Responsibility Operational

当系统给出错误或越权回答时，首先需要限制影响，再追查原因。可以暂停受影响索引或查询路径，保留必要的请求与版本记录，并通知负责数据与服务的人员。调查应沿证据链定位：源数据是否错误，观察是否抽取失真，实体是否误合并，权限是否遗漏，生成是否超出支持范围。不同原因需要不同修复，不能把所有事故都归结为语言模型幻觉。恢复服务前，应使用原始失败案例及邻近变体确认问题已经解决。

责任分工应在上线前确定。数据所有者负责来源与适用范围，知识维护者负责本体与合并规则，服务团队负责检索、权限和可用性，业务负责人决定哪些回答可以影响行动。用户纠错与申诉也需要明确入口。事故报告可以记录影响、根因、修复、验证和防复发措施，并将敏感细节限制在

必要范围。衡量成熟度时，不只看从未出错的宣传，而应看发现速度、修复速度与复发控制。图谱的可追踪性在这里产生实际价值：它能把一段错误答案拆解成可负责、可修改的环节，并帮助组织用证据改进流程。

When a system produces an incorrect or unauthorized answer, first limit the impact and then investigate. Affected indexes or routes may be paused while necessary request and version records are preserved and responsible staff are informed. Investigation follows the evidence chain: incorrect sources, distorted extraction, mistaken merges, missed permissions, or generation beyond support. Different causes need different repairs; calling every incident hallucination prevents diagnosis. Before restoration, the original failure and nearby variants should confirm the fix.

Responsibilities should be defined before deployment. Data owners govern sources and scope, knowledge maintainers ontologies and merges, service teams retrieval, permissions, and availability, and business owners decide which outputs may influence action. Corrections and appeals need clear entry points. Incident records can cover impact, cause, repair, verification, and prevention while restricting sensitive detail. Maturity concerns detection speed, repair speed, and recurrence control rather than claims of never failing. Traceability makes a practical contribution by decomposing a wrong answer into accountable and editable stages and helping organizations improve through evidence.

返回目录 Back to contents

第12章　企业与科研应用蓝图

Enterprise and Scientific Application Blueprints

设备运维中的事件证据链

Event Evidence Chains for Equipment Maintenance

以下场景是用于讨论的应用蓝图，而非已经完成部署的案例。设备运维可将设备、部件、故障事件、工单和维护手册连接起来，图片用于记录现场状态，文字用于说明操作与更换。查询“这台设备反复出现同类告警的相关记录有哪些”时，系统先确认设备身份和时间范围，再查找告警、工单与部件更换记录。图的作用是连接分散证据，不能仅凭连续发生就认定因果。答案应区分已记录原因、工程师假设和仍待排查的关联。

该蓝图的关键是事件建模。一次告警不等于一个永久属性，同样的错误代码也可能由不同原因触发。图片必须关联拍摄时间与设备编号，避免把另一台相似设备当作当前对象。评测可选择历史问题，要求系统找到完整记录并保留未解决状态，再由工程人员判断是否减少查找时间。错误类型包括跨设备误合并、遗漏维修后的新版本以及把暂时现象当作长期故障。系统适合辅助检索和准备检查材料，实际维修动作仍应按既定专业流程执行。收益应通过任务耗时、证据遗漏和错误严重性测量，而不只看答案是否像专家写的。

The following scenarios are proposed blueprints rather than completed deployments. Maintenance can connect equipment, components, incidents, work orders, and manuals. Images document site conditions

and text records operations and replacements. A question about repeated alarms first requires device identity and temporal scope, followed by related alarms, work orders, and component changes. The graph connects scattered evidence but cannot infer causation from sequence alone. Answers should distinguish recorded causes, engineer hypotheses, and unresolved associations.

Event modeling is central. An alarm is not a permanent property, and one error code can have different causes. Images need capture time and equipment identifiers to avoid confusing similar machines. Historical tasks can test complete record recovery and preservation of unresolved status, with engineers judging search time savings. Errors include merging devices, missing post repair versions, and treating temporary conditions as persistent faults. The system supports retrieval and preparation of inspection material; maintenance actions follow established professional procedures. Benefits should be measured through task time, omitted evidence, and error severity rather than expert sounding prose alone.

产品服务与版本化知识

Product Support and Versioned Knowledge

产品服务常见问题不是完全没有资料，而是资料适用于不同型号、地区和软件版本。图谱可以表达产品族、具体型号、部件、功能、发布日期与适用规则，文档页面和截图成为证据。用户问某功能是否支持时，系统应先核实版本，再定位相关手册和变更记录。如果当前版本缺少确认资料，应返回不确定，而不能从相邻型号或旧截图推断支持。多模态信息在这里的价值包括识别界面位置和解释操作步骤，但截图同样需要版本条件。

评测可以构造同名功能跨版本变化的成对问题，检查系统是否把正确资料匹配到正确对象。也可以加入过时文档、地区限定和仅适用于特定配置的例外，观察这些条件是否进入答案。服务人员应能快速查看引用页面，而不是在长篇回答中再次搜索。图谱结构有助于定位兼容关系，但兼容不应自动传递：产品 A 支持部件 B，部件 B 与产品 C 相关，并不能推出 A 支持 C。业务规则应以明确类型与适用范围表示。这个应用的核心收益是减少错误适配和重复查找，其验证需要真实服务任务，而不是只评估通用问答语言质量。

Product support often has abundant material applying to different models, regions, and software versions. A graph can represent families, models, components, features, release dates, and applicability rules, with pages and screenshots as evidence. A feature question should establish the version before locating manuals and changes. Missing confirmation should produce uncertainty rather than inference from neighboring models or old screenshots. Visual material can identify interface locations and explain steps, but screenshots need version conditions too.

Paired questions about a feature changing across versions can test correct evidence assignment. Outdated documents, regional restrictions, and configuration exceptions test whether qualifiers reach the answer. Support staff should be able to inspect cited pages quickly. Graph structure helps locate compatibility, but compatibility is not automatically transitive. A product supporting a component and that component being related to another product do not establish mutual support. Rules need explicit types and scope. The main potential benefits are fewer inappropriate matches and less repeated searching, to be validated through actual support tasks rather than generic language quality.

科研文献中的主张方法与证据

Claims Methods and Evidence in Scientific Literature

科研场景中，论文、数据集、方法、任务和实验结果可以形成相互关联的知识网络。一个问题可能要求比较两种方法在不同数据条件下的表现，单纯检索最相近摘要容易忽略设置差异。图谱可以把结果连接到数据划分、指标、模型与论文版本，帮助研究者找到可比较的证据。图表页面提供数值上下文，正文提供限定条件，补充材料可能提供关键实现细节。系统应区分作者报告的结果、其他论文复现的结果和研究者自己的推断，不能把所有边都写成已证实结论。

该应用尤其需要版本和发表状态。预印本更新可能修改实验，会议版本与期刊扩展版可能属于同一研究系列；若重复计数，会夸大证据数量。引用网络的中心性也不等于方法正确性，常被引用可能反映基础性、争议性或历史原因。评测可要求系统回答一个带条件的比较问题，并逐条列出可比与不可比之处。人工核验应关注实验单位、指标方向与是否存在未报告结果。检索工具适合提高材料发现效率，研究判断仍需要对原文进行批判性阅读。本报告使用 Consensus 与 SciSpace 进行文献发现，也采用这一分工。

Scientific literature can connect papers, datasets, methods, tasks, and experimental results. Comparing methods under different data conditions requires more than similar abstracts. A graph can link results to splits, metrics, models, and paper versions, helping researchers identify comparable evidence. Figure pages provide numerical context, text supplies qualifiers, and supplements may contain implementation details. Author reported results, independent reproductions, and a researcher's inferences should remain distinct rather than becoming equally confirmed edges.

Version and publication status are especially important. Preprint updates may revise experiments, while conference and extended journal papers can belong to one research series. Counting them independently exaggerates evidence. Citation centrality is not correctness: frequent citation may reflect foundational status, controversy, or history. Evaluation can ask a conditional comparison and require explicit comparable and incomparable aspects. Human review should inspect experimental units, metric directions, and missing results. Discovery tools improve access to material, while research judgment still requires critical reading of originals. This report follows that division when using Consensus and SciSpace for literature discovery.

组织知识中的主题发现与少数观点

Theme Discovery and Minority Views in Organizational Knowledge

会议记录、项目复盘和内部报告往往包含跨文档主题，适合讨论全局 GraphRAG 的使用条件。用户可能问“过去一段时间反复出现哪些协作障碍”，系统需要覆盖多个项目，而不是只找到一篇措辞相近的记录。社区报告可以作为主题入口，但最终归纳仍应追溯到具体来源，并区分反复出现、偶尔提到和证据不足。不同团队对同一事件的解释可能相互冲突，系统不应为了生成整齐主题而抹去差异。

评估这种应用可以邀请熟悉资料的人员先建立主题清单，再检查系统是否发现重要主题、是否遗漏少数观点、是否把个别事件扩大为组织规律。应控制文档数量不均造成的偏差：发言多的团队

不一定更有代表性，重复周报也不代表独立事件增加。可以报告主题对应的独立项目数与时间范围，而不是只给提及次数。隐私与权限尤其重要，因为跨文档组合可能暴露个人关系。适合面向管理的问题也应避免对个人作无证据归因。图增强发现的价值在于帮助提出可核验问题和整理线索，而不是用自动摘要代替组织调查。

Meeting notes, retrospectives, and internal reports often contain themes spanning documents. A question about recurring collaboration barriers requires multiple projects rather than one similarly worded note. Community reports can provide entry points, but synthesis should trace to sources and distinguish repetition, occasional mention, and insufficient evidence. Teams may interpret one event differently, and tidy themes should not erase disagreements.

Evaluation can ask knowledgeable reviewers to establish themes independently, then assess important discoveries, minority omissions, and overgeneralization from isolated incidents. Unequal document volume needs control: more talkative teams are not necessarily more representative, and repeated weekly reports do not create independent events. Theme summaries can report independent projects and temporal scope instead of only mentions. Permissions are critical because cross document combinations may reveal personal relationships. Management questions should also avoid unsupported attribution to individuals. The value lies in organizing clues and generating verifiable questions rather than replacing organizational investigation with automatic summaries.

供应与商品关系中的匹配边界

Matching Boundaries in Supply and Product Relations

商品与供应资料经常包含相同名称、不同规格，以及不同名称、实际相同的对象。多模态图谱可结合包装图片、文字规格、商品编号和供应记录建立候选匹配，再用明确规则确认。需要区分同一商品、同类商品、替代商品和配套商品，这些关系对业务动作的含义不同。图片相似可以帮助发现候选，但不能独立决定库存合并或供应替换。商品版本、包装数量和计量单位应成为身份判断的一部分。

评测应加入困难负例，例如同品牌不同容量、外观相似但型号不同，以及改版包装对应同一商品。匹配错误的代价通常不对称：错误合并可能污染库存与分析，漏匹配则增加人工工作。阈值应根据这种代价选择，而不是一味追求最高总体准确率。审核界面可以展示关键差异及原始证据，并允许保留待判定状态。图结构还可用于发现异常，例如一个编号突然连接到不兼容规格，但异常只是核查线索。应用是否成功，应看真实匹配任务的误合并、漏合并、审核时间与后续纠错，而非只看一组漂亮的商品关系网络。

Supply and product records contain identical names with different specifications and different names for the same object. Multimodal graphs can combine packaging, specifications, identifiers, and supply records to propose matches, followed by explicit confirmation rules. Same product, same category, substitute, and accessory are different relations with different business consequences. Visual similarity helps discover candidates but cannot alone authorize stock merging or supply substitution. Versions, package quantities, and measurement units belong in identity decisions.

Hard negatives include one brand with different capacities, similar looking models, and redesigned packaging for the same product. Error costs are asymmetric: false merges can contaminate inventory and

analysis, while missed matches create manual work. Thresholds should follow those costs rather than maximum aggregate accuracy alone. Review interfaces can show decisive differences and source evidence and preserve unresolved status. Graph structure may flag anomalies such as one identifier linked to incompatible specifications, but anomalies remain inspection clues. Success concerns false and missed merges, review time, and later corrections, not the appearance of an attractive network.

共用评估框架与场景差异

A Shared Evaluation Framework with Domain Specific Conditions

上述应用共享一条证据流程：明确对象与时间，定位多源材料，建立关系，保留限定条件，再生成可核验输出。但它们的正确性标准不同。设备场景关注事件与专业流程，产品场景关注版本兼容，科研场景关注实验可比性，组织场景关注代表性，商品场景关注身份匹配。把所有场景放进同一个通用问答基准，容易忽略真正重要的错误。因此公共平台可以共享资源管理、证据接口和审计能力，领域层则需要独立本体、规则和评测集。

实施时可先选一个高频任务建立完整闭环，再判断哪些能力值得推广。跨部门复用不仅是复制代码，还要检查术语、权限、源数据责任和人工审核能力是否一致。某个部门成功使用图谱，不代表其他部门的数据已经具备同样条件。报告建议为每个场景保留一页决策卡：目标用户、问题分布、必要证据、关键错误、更新频率、责任人和停止条件。这样的卡片让研究概念转化为可管理的实验，也使不同场景之间的比较有共同语言。最终价值来自具体任务中更少的遗漏和更好的判断，而不是采用了多少种模型或数据库。

These applications share an evidence workflow: establish objects and time, locate sources, connect relations, preserve qualifiers, and generate verifiable outputs. Their correctness criteria differ. Maintenance emphasizes events and professional procedures; support version compatibility; research experimental comparability; organizations representativeness; and products identity. One generic benchmark can miss decisive errors. A common platform may share resource management, evidence interfaces, and auditing, while domains need their own ontologies, rules, and test sets.

Implementation should close the loop for one frequent task before expanding. Cross departmental reuse requires checking terminology, permissions, source ownership, and review capacity, not just copying code. Success in one department does not establish equivalent data conditions elsewhere. A scenario decision card can record users, question distributions, required evidence, critical errors, update frequency, owners, and stopping conditions. Such cards turn concepts into manageable experiments and provide a common language across domains. Value comes from fewer omissions and better decisions in concrete tasks rather than the number of models or databases adopted.



图 19 同一平台需要不同场景的验收标准 / Shared platform, domain specific acceptance

来源 / Source 本报告原创设计 / Original design

返回目录 Back to contents

第13章　教育文化与视觉思维实践

Education Culture and Visual Thinking Practice

先让学生看懂证据如何支持结论

Teach How Evidence Supports a Conclusion

面向人工智能通识教学，可以先从一小组资料开始，让学生画出“来源到陈述到结论”的关系，而不是直接要求搭建复杂系统。每张证据卡只记录一个可核验陈述，附来源、位置、时间与状态；结论卡必须连接到足够证据，并标明哪些连接是推断。视觉思维在这里不是装饰，而是把支持关系外显。学生能够看到一句流畅回答可能依赖多份资料，也能看到一个没有来源的图节点并不自动成为事实。

课堂可以采用先个人、再小组、最后全班比较的流程。个人标注资料中的对象和条件，小组协商实体是否相同，全班比较不同图谱为何产生不同答案。教师不必追求唯一漂亮图形，应重点讨论边的语义是否明确、否定是否保留、证据能否定位。为了降低工具门槛，可以先用纸卡或普通演示软件，随后再引入图数据库和检索程序。学习评价关注证据组织和错误解释，而不只看节点数量。一个节点较少但出处清楚、能表达未知的图，往往比大而混乱的自动图更能体现学生理解。

Introductory AI teaching can begin with a small source collection and ask students to map sources to claims and conclusions before building a complex system. Each evidence card records one verifiable statement with origin, location, time, and status. Conclusion cards need sufficient support and explicit inference links. Visual thinking makes support relations visible rather than decorative. Students see that fluent answers may depend on several records and that an unsourced node does not automatically become a fact.

A lesson can progress from individual annotation to group agreement and class comparison. Individuals identify objects and conditions, groups negotiate identity, and the class examines why different graphs produce different answers. The goal is not one attractive layout but clear edge meanings, preserved negation, and locatable evidence. Paper cards or ordinary presentation software can precede graph databases and retrieval code. Assessment should emphasize evidence organization and error explanation rather than node count. A smaller graph with clear provenance and explicit unknowns can demonstrate more understanding than a large disordered automatic graph.

用对照活动理解 RAG 与 GraphRAG

Use Controlled Activities to Understand RAG and GraphRAG

课堂对照可以让同一组学生用三种方式解决同一问题：只凭已有知识回答，查找独立片段回答，利用关系卡组织证据后回答。教师准备直接事实、跨片段连接、全局主题和无答案问题各一例，要求学生记录找到什么、遗漏什么以及引用来自哪里。这样能把抽象方法差异转化为观察到的过程。对照的目的不是安排 GraphRAG 必胜，而是让学生发现不同问题对证据结构的要求不同。

为了避免不公平比较，三组应使用相同资料与时间预算，并轮换方法。结果讨论可以区分正确答案、有效引用、必要证据完整性和过度推断。对于跨片段问题，关系卡可能帮助发现连接；对于简单事实，额外画图可能只增加时间；对于无答案问题，任何方法都应承认缺口。学生还可以故意加入一条错误边，观察答案如何改变。这一活动强调图结构既能帮助推理，也能传播错误。最后要求每组写一段带条件的结论，例如“在这组资料的关系问题中有帮助”，而不是“图方法一定更好”。这种表述训练将实验观察与普遍论断分开。

A controlled classroom activity can solve one question through three routes: prior knowledge alone, independent passage lookup, and relationship cards organizing the same evidence. The teacher supplies a factual, relational, global thematic, and unanswerable question, asking students to record discoveries, omissions, and citation locations. Abstract method differences become observable processes. The activity should not guarantee a GraphRAG victory; it should reveal that questions require different evidence structures.

Groups need equal material and time and should rotate methods. Discussion separates answer correctness, citation support, necessary evidence coverage, and overinference. Relationship cards may help a cross passage question, while drawing a graph for a simple fact may only add time. Every method should acknowledge an unanswerable question. Students can insert an erroneous edge and observe its effect, showing that structure can propagate errors as well as assist reasoning. Each group then writes a conditional conclusion, such as usefulness for relational questions in this source set, rather than claiming universal superiority. This trains the distinction between an observation and a general law.

图文对齐活动训练视觉判断

Image Text Alignment Activities Train Visual Judgment

多模态活动可以提供若干菜品文字、图片和菜单版本，其中包含正确配对、相似外观、过期图片和无法判断的材料。学生先独立给出候选，再列出支持身份匹配的证据。教师要求区分看到的现象、推测的类别和已确认属性。例如，看见红色小块是一种观察，把它识别为辣椒是一个候选判断，认定某商品含辣椒还需要身份和版本依据。通过分层表达，学生能理解视觉模型为何不能直接承担全部事实确认。

活动图形可以用三种边型表达不同语义：资源指向观察、观察指向候选实体、确认记录支持属性。颜色只辅助分类，边旁文字承担主要含义，避免色觉差异或黑白打印使图失去可读性。还可让学生裁剪图片并交换给另一组，检查丢失上下文是否改变判断。若去掉菜单日期或图例，原先看似确定的配对可能不再成立。评价时奖励合理拒绝和补证建议，而不是强迫每张图都有唯一答案。视觉思维的学习目标，是让学生看清证据边界，而不是把所有信息变成更密集的图片。

A multimodal activity can provide dish descriptions, images, and menu versions containing correct pairs, similar appearances, outdated photographs, and unresolved cases. Students propose candidates and list identity evidence. They distinguish observed features, inferred categories, and confirmed attributes. Seeing small red pieces is an observation; identifying them as chilli is a candidate judgment; asserting chilli in a specific product also requires identity and version evidence. Layered expression explains why visual models cannot perform all factual confirmation directly.

Three edge types can connect resources to observations, observations to candidate entities, and confirmations to attributes. Colour supports categories while written edge labels carry meaning, preserving readability for colour vision differences and monochrome printing. Students can crop images for another group and test whether lost context changes judgment. Removing a menu date or legend may undermine an apparently certain match. Assessment should reward justified rejection and requests for evidence rather than force one answer for every image. The goal of visual thinking is to expose evidence boundaries, not merely produce denser pictures.

文化资料中的多种解释与出处

Multiple Interpretations and Provenance in Cultural Collections

文化与历史资料经常包含异名、版本、传说和学术争议。多模态图谱可以连接作品、人物、地点、图像局部和文献记载，但必须区分历史事实、作者解释和后世传说。某张图像被描述为表现一个故事，不等于故事中的事件已被历史证据确认。系统可以保存多种解释及各自出处，而不是为每个对象只保留一个标准答案。图谱在这里的价值是帮助比较证据和解释路径，而非消除所有争议。

一个课程项目可以选择少量公开文化资料，建立作品版本与解释关系。学生需说明图片来自哪个收藏或出版版本、裁剪区域是什么、解释依据在哪里，并记录不确定年代或归属。中英术语也要谨慎，有些文化概念没有完全等价的单词，可保留原文名称并提供解释。评价可围绕出处准确、解释区分、图像定位和对争议的公平呈现。若用于面向公众的展示，还要核实图片使用条件与署名要求。这里的研究设计关注证据组织，不自动授权复制或发布所有检索到的图像。文化理解的丰富性可以被图谱呈现，而不必被压缩成单一类别。

Cultural and historical sources contain alternate names, editions, legends, and scholarly disputes. A multimodal graph can connect works, people, places, image regions, and records while distinguishing historical claims, interpretations, and later traditions. An image interpreted as depicting a story does not confirm that the story's events occurred. Multiple interpretations can retain separate sources instead of assigning each object one standard answer. The graph helps compare evidence and interpretive paths rather than eliminate disagreement.

A course project can use a small public collection to connect editions and interpretations. Students identify image collections or publication versions, cropped regions, supporting texts, and uncertain dates or attributions. Bilingual terminology needs care because some cultural concepts lack exact equivalents; original terms with explanations may be preferable. Assessment can examine provenance, separation of interpretation, visual localization, and fair treatment of disputes. Public displays also require checking image conditions and attribution. The research design concerns evidence organization and does not automatically authorize copying or publishing every retrieved image. Graphs can represent cultural richness without reducing it to one category.

实操作业与可复核的学习评价

Practical Assignments and Verifiable Learning Assessment

作业可以要求学生提交一个小型证据图、两类问题的回答、一份错误分析和一段方法反思。资料规模应足够让人检查，而不以爬取数量作为成绩。证据图必须包含至少一处时间条件、一处不确定观察和一处可能冲突；回答必须连接到来源位置。学生可以使用模型辅助抽取，但需记录模型候选如何被审核。这样的要求把注意力从自动生成结果转向可解释过程，也便于教师判断学生是否理解概念而非只会调用工具。

评分维度可分为对象与关系定义、证据追踪、回答边界、纠错质量和表达清晰度。每个维度采用明确描述，例如“引用能定位但不能充分支持”与“引用直接支持且条件完整”应属于不同等级。学生互评可先用同一示例校准，再评对方作品，并要求指出一处具体可改进位置。教师抽查高分和争议样本，防止只按视觉美观评分。反思部分应回答图结构帮助了什么、增加了什么成本、在哪些问题上没有帮助。评价标准因此同时承认成功与失败的学习价值。一个诚实发现图并无增益的实验，也可能比无证据宣称效果更好的作业更优秀。

An assignment can require a small evidence graph, answers to two question types, an error analysis, and a methodological reflection. The collection should be inspectable rather than graded by crawling volume. Graphs should contain a temporal condition, an uncertain observation, and a potential conflict; answers should link to source locations. Models may assist extraction if students record how candidates were reviewed. These requirements emphasize an explainable process and help teachers distinguish understanding from tool operation.

Rubrics can assess object and relation definitions, traceability, answer boundaries, correction quality, and clarity. Descriptions should distinguish a locatable but insufficient citation from direct support preserving all conditions. Peer reviewers can calibrate on one example before reviewing another project and identifying a specific improvement. Teachers can sample high scoring and disputed work to avoid grading visual polish alone. Reflection should explain where structure helped, what it cost, and where it did not help. This recognizes learning from both success and failure. An honest experiment finding no graph benefit may be stronger than an unsupported claim of improvement.



图 20 课堂学习以证据组织与纠错为中心 / Learning through evidence and correction
来源 / Source 本报告原创设计 / Original design

双语教学与认知负荷控制

Bilingual Teaching and Cognitive Load

中英对照有助于连接中文课堂与英文研究资料，但逐字堆叠两种语言会增加阅读负担。建议以中文提出问题与解释关键边界，英文保留对应概念和研究表达，并通过同一图形保持语义一致。术语表应说明常见误解，例如 grounding 强调与依据连接，faithfulness 强调对给定证据的忠实，二者都不能单独保证现实事实正确。图中的标签宜短，长解释放在正文或讲解中。双语的目的在于促进概念迁移，而不是让每张图承载两篇文章。

教学节奏可以采用问题、图示、操作、讨论和迁移五个阶段，每阶段只引入少量新概念。先让学生经历查找困难，再说明图结构的作用，通常比先展示完整复杂架构更容易理解。图示应保持对象颜色和边语义稳定，减少无意义动画与装饰。对于不同背景学生，可提供纸卡、表格和代码三种入口，但使用同一证据任务和评价标准。最后让学生把方法迁移到自己专业中的一个小问题，并说明哪些条件改变了。这样课程既保留技术严谨性，也避免把 GraphRAG 讲成只有程序员能够理解的黑箱，或讲成可以解决所有知识问题的通用口号。

Bilingual teaching connects Chinese classes with English research, but stacking both languages word for word raises reading load. Chinese can frame questions and boundaries while English supplies corresponding concepts and research expression, with shared diagrams maintaining meaning. A glossary should address common confusions: grounding concerns connection to evidence, faithfulness concerns adherence to supplied evidence, and neither alone guarantees factual correctness. Figure labels should remain short, with longer explanations in prose or narration. The purpose is conceptual transfer rather than placing two essays inside every diagram.

A lesson can move through question, diagram, activity, discussion, and transfer, introducing few concepts at each stage. Experiencing retrieval difficulty before seeing graph structure can be more accessible than starting with a complete architecture. Stable colours and edge meanings reduce unnecessary cognitive effort. Paper, table, and code entry routes can serve different backgrounds while sharing tasks and assessment. Students finally transfer the method to a small disciplinary problem and explain changed

conditions. This preserves rigor without presenting GraphRAG as either a programmer only black box or a slogan capable of solving every knowledge problem.

返回目录 Back to contents

第14章　校园菜品多模态 GraphRAG 应用设计

A Campus Dish Multimodal GraphRAG Blueprint

把需求限定为信息查找与核验

Bound the Task to Information Discovery and Verification

校园菜品系统可以作为连接两篇核心文章的具体设计对象。目标不是自动决定学生应该吃什么，而是帮助查找某窗口某版本菜品的已知信息、相关图片和待确认事项。用户可以问“哪些记录提到鸡肉”“这张图可能对应哪道菜”“两个同名商品是否属于同一版本”。这些问题既涉及对象与关系，也涉及图文身份与证据。系统输出应包括结论、来源、时间以及仍未知的部分，避免把常见配方当成当前食堂的实际配料表。

需求阶段可访谈学生与管理人员，收集真实问题并分类：精确商品查询、食材相关查找、图片匹配、反馈主题和信息冲突。不同类别设置不同验收标准。对无法由现有资料可靠回答的问题，应标明需要窗口确认或补充当前版本资料。范围越清楚，越容易选择适当技术：并非每个问题都需要全局社区分析，也并非每次都要调用视觉模型。一个有价值的最小系统，可以只处理少量窗口与明确日期，在数据更新和纠错上形成闭环。报告后文的模拟资料用于验证这种信息逻辑，不代表任何真实校园供餐情况。

A campus dish system offers a concrete design connecting the two anchor articles. Its purpose is to find known information, related images, and unresolved issues for a stall's dish version, rather than decide what a student should eat. Questions may concern records mentioning chicken, possible matches for an image, or whether same named products share a version. They require relations, visual identity, and evidence. Outputs should state sources, time, and unknowns instead of treating a common recipe as an actual canteen ingredient list.

Requirements can collect real student and staff questions and classify precise lookup, ingredient discovery, image matching, feedback themes, and conflicts. Each needs acceptance criteria. Questions unsupported by records should request current information or stall confirmation. Clear scope makes technology selection easier: global communities and visual models are not needed for every request. A useful minimum system can cover a few stalls on specified dates while closing update and correction loops. The later simulated records test information logic and describe no actual campus food service.

建模菜品版本与观察来源

Model Dish Versions and Observation Sources

建议区分菜品概念、窗口商品和商品版本。宫保鸡丁作为概念可以连接常见做法，窗口 A 的宫保鸡丁是具体商品，某日期生效的配方则是版本。食材关系应优先连接到版本，而不是无限期连接到菜名。菜单、窗口确认、图片和评论作为不同来源产生观察；图片中疑似食材不直接覆盖已确认配料。评论中的口味反馈也应连接到体验事件，避免一条评价变成所有时间都成立的属性。

这个模型能表达变化与不一致。例如，一次配方调整后，旧图片仍可能出现在页面上；系统应发现图片与当前版本的时间不匹配，而不是把两者合并。若只知道拍摄时间而不知道适用版本，可以保留待关联状态。菜单名称变化不一定代表商品变化，同名也不保证同一商品，因此身份判断需要窗口、时间和管理记录。数据维护界面应允许工作人员更新版本、确认属性和撤回错误图片。图谱面向用户的视图可以简化显示，但底层证据关系应完整。这样用户看到的是清楚的当前信息，维护者仍能追踪它如何形成。

The model should distinguish a dish concept, a stall product, and a product version. Kung pao chicken as a concept may connect to common preparation, the item sold by stall A is a product, and its recipe effective on a date is a version. Ingredient relations should attach primarily to versions rather than permanently to names. Menus, staff confirmations, photographs, and reviews produce different observations. Suspected visible ingredients should not overwrite confirmed lists, and taste reviews should connect to experience events instead of becoming timeless properties.

The model handles change and disagreement. An old photograph remaining after a recipe update creates a temporal mismatch rather than a reason to merge evidence. A known capture date without a known version can remain unresolved. Renaming does not necessarily change identity, and identical names do not guarantee identity, so stall, time, and management records matter. Staff interfaces should support version updates, attribute confirmation, and image withdrawal. User views may simplify presentation while underlying evidence remains complete. Users receive clear current information and maintainers retain its history.

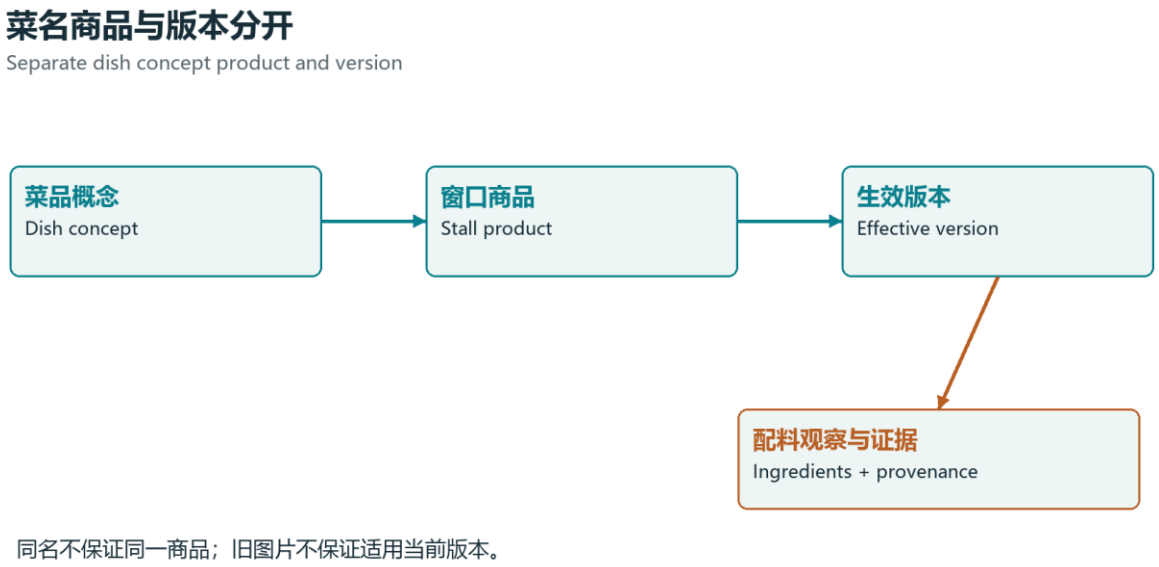


图 21 版本化菜品模型 / Versioned dish model
来源 / Source 本报告原创设计 / Original design

文字图像与人工确认的融合流程

Fuse Text Images and Human Confirmation

接入菜单后，文本抽取器识别菜名、食材和做法，并保留原文位置。图片处理先检查清晰度与商品关联，再生成可能的食材类别与区域。融合阶段比较商品身份、版本和语义条件，把一致信息作为支持，把不一致信息交给核验队列。对不可见调味料，图片不提供排除证据；对文字中的否定，抽取器必须保留否定状态。人工确认记录应独立保存，并说明确认范围，例如只确认公开配料描述，而不声称验证所有加工过程。

流程可以采用低风险自动通过与高影响人工核验相结合。明确商品编号和当前菜单之间的匹配可以按规则处理，模糊图片或冲突配料则保留候选。审核界面并排展示菜单文字、图片区域和旧版本信息，减少来回查找。每次确认后生成可查询视图，同时保留原始候选以便研究错误。质量抽样应覆盖不同窗口、菜类与拍摄条件，而不只抽最清楚图片。这个流程的重点是把多模态补充转化为可管理证据，而不是把视觉识别分数直接当作可信程度。模型帮助发现线索，确认规则决定哪些线索进入当前知识。

Menu extraction identifies names, ingredients, and preparation while preserving source locations. Image processing first checks quality and product association, then proposes ingredient categories and regions. Fusion compares identity, version, and semantic conditions, treating agreement as support and disagreement as a review item. Images cannot establish the absence of invisible seasoning, and text extraction must preserve negation. Human confirmation should remain a separate record with explicit scope, such as confirming a published ingredient description without claiming verification of every processing step.

Rules can automate low risk matches while high impact ambiguity receives review. Explicit product identifiers may match current menus automatically, while unclear images and conflicting ingredients remain candidates. A review interface can show text, regions, and older versions together. Confirmations generate queryable views while preserving original candidates for error analysis. Sampling should cover stalls, dish categories, and camera conditions rather than only clear photographs. The goal is manageable evidence rather than converting a detector score directly into trust. Models discover clues; confirmation rules determine which clues enter current knowledge.

为不同问题选择检索路径

Choose Retrieval Routes for Different Questions

精确商品问题先查当前版本与已确认属性，再返回菜单或确认记录。食材查找从食材实体反向连接相关商品，同时区分已确认、仅文字提及和视觉候选。图片匹配先产生候选，再检查窗口与时间；若用户没有提供这些条件，结果可以是候选列表而非唯一身份。反馈主题问题可以汇总评论事件，但需要说明样本时间和来源覆盖，不把少量评论概括为全体学生意见。这些路径可以共享证据标识，却不必使用完全相同的检索算法。

答案界面应让用户看到信息状态。例如“记录中有鸡肉”与“确认不含某食材”是不同陈述，后者需要明确排除证据。对于冲突记录，可以列出双方来源与时间，并提示尚待确认。系统应避免把未检索到关系显示为无该属性，也应避免把评论中的个人感受变成客观商品标签。若问题需要全局比较，可以先说明可比较的窗口与日期，再生成结论；缺乏同条件数据时不强行排名。这个设计将 GraphRAG 的关系组织能力用于缩小信息差，而不是制造超出资料支持范围的确定答案。

Precise product questions first retrieve current versions and confirmed attributes, then return menu or confirmation records. Ingredient discovery traverses from ingredients to products while separating confirmation, textual mention, and visual candidates. Image matching generates candidates before checking stall and time; missing context can lead to a candidate list rather than one identity. Feedback themes can summarize review events but must state sampling periods and coverage instead of treating a few comments as all students' views. Routes share evidence identifiers without requiring identical algorithms.

Interfaces should expose information status. A record mentioning chicken differs from confirmed absence of an ingredient, which requires explicit exclusion evidence. Conflicts can show both sources and dates pending confirmation. Missing retrieved edges should not become absent attributes, and personal taste reports should not become objective product facts. Global comparisons should first define comparable stalls and dates. Without matched information, rankings should remain limited. This applies graph organization to information gaps without generating certainty beyond the sources.

数据运营与反馈主题分析

Data Operations and Feedback Theme Analysis

校园应用需要持续维护菜单与图片，不只是一次构图。可以设置每日或按变更更新的资源清单，由窗口人员确认当前商品版本，学生反馈进入待审核队列。更新频率应取决于真实变化，不必为

追求实时而持续重复处理不变文件。反馈分类可以区分信息错误、图片不匹配、版本过期和体验意见，分别交给合适责任人。系统应记录从发现到修复的时间，让维护能力成为可测指标。

主题分析可以帮助发现重复问题，例如某些窗口图片长期过期或菜单描述缺少关键属性，但需要防止重复提交放大声量。可按独立事件和时间聚合，并保留少数不同意见。对于经营分析，系统只能描述资料中的模式，不能仅凭评论共现断言经营原因。若某窗口晚高峰订单多且等待久，可以提出待核查关联，而不是直接认定订单量导致全部延迟。改善建议还需要结合排班、设备和现场流程等资料。多源图谱的作用是把这些核查路径组织起来，帮助管理人员提出更准确的问题，而不是自动替代现场判断。

A campus application requires continuing menu and image maintenance, not one graph build. Resource manifests can update daily or when changes occur, with staff confirming current versions and student corrections entering review. Frequency should follow actual change rather than repeatedly processing unchanged files for apparent real time operation. Feedback categories can separate incorrect information, mismatched images, outdated versions, and experiences, routing each to responsible people. Time from discovery to correction becomes a maintenance measure.

Theme analysis can reveal recurring issues such as stale images or missing attributes, while repeated submissions should not amplify one event. Aggregate independent events over time and preserve minority views. Operational analysis describes patterns in records but cannot infer causes from comment cooccurrence. High evening order volume and long waits suggest an association to investigate, not proof that volume explains every delay. Improvement decisions require staffing, equipment, and process evidence. The graph organizes these investigation routes and helps staff ask better questions rather than replacing observation and judgment.

试点验收与扩展条件

Pilot Acceptance and Conditions for Expansion

试点可选择少量窗口、固定日期范围和一组高频问题。验收同时检查信息正确、证据可追踪、冲突不被隐藏和更新能够生效。可以让工作人员与学生分别完成查找任务，记录耗时、错误和是否理解回答中的不确定性。若系统只提高查找速度却让用户误以为所有属性都已确认，仍需改进界面。无答案问题也必须纳入测试，确保系统能引导补充资料。最终评估应把自动指标、人工任务结果和维护负担放在一起。

扩展到更多窗口之前，应确认身份规则、本体和更新流程是否仍适用。不同商家可能使用不同命名方式，图片质量和配料公开程度也不同。可以先增加一组差异较大的窗口作为迁移测试，而不是只增加相似数据。若质量下降，应定位是数据不足、身份歧义还是流程无法维持。报告建议设定停止条件：关键错误无法及时修复、版本长期失效或证据来源不足时，应缩小回答范围。一个可信应用可以在有限范围内稳定工作，并清楚表达边界。扩展的依据应是新范围中的实测表现，而不是已有图谱节点数量达到某个规模。

A pilot can cover a few stalls, fixed dates, and frequent questions. Acceptance checks correctness, traceability, visible conflicts, and effective updates. Staff and students can perform search tasks while measuring time, errors, and understanding of uncertainty. Faster lookup is insufficient if the interface

makes every attribute appear confirmed. Unanswerable cases should test guidance toward missing evidence. Automated measures, human task outcomes, and maintenance effort belong in one assessment.

Before expansion, verify that identity rules, ontology, and update processes still apply. Vendors differ in naming, image quality, and ingredient disclosure. Add deliberately different stalls as a transfer test rather than only similar data. Quality declines should be traced to insufficient evidence, ambiguous identity, or unsustainable processes. Stopping conditions can include unresolved critical errors, persistently stale versions, or inadequate sources, leading to narrower answer scope. A trustworthy application works consistently within a defined range and expresses its boundaries. Expansion should follow measured performance in the new range rather than the number of graph nodes already accumulated.

返回目录 Back to contents

第15章　可复算的微型图谱实验

A Reproducible Miniature Graph Experiment

资料与实验范围

Records and Experimental Scope

本实验使用七条人工编写资料，不涉及真实门店或真实供餐。D1 描述窗口 A 宫保鸡丁含鸡肉和花生；D2 描述窗口 B 麻婆豆腐含豆腐和牛肉；D3 描述窗口 A 晚高峰等餐较久；D4 描述同一窗口晚高峰订单较多；D5 描述窗口 C 鸡肉沙拉含鸡肉和生菜；D6 描述该沙拉口味清淡；D7 描述窗口 A 宫保鸡丁不含花生。为单独演示冲突，实验暂把 D1 与 D7 视为适用于同一版本；真实系统必须先核实这个条件。

实验中的关系由人工明确录入，程序只进行确定性筛选、状态判断和算术计算，没有调用语言模型抽取、视觉模型识别或真实 GraphRAG 框架。因此结果能够证明代码遵循了给定逻辑，却不能证明自动构图、语义检索或生成质量。这个范围说明是实验设计的一部分。资料规模很小，便于读者逐条检查，也适合课堂手工复算。后续若接入模型，应把人工关系作为参考标注，并另行测试抽取错误、同名实体和版本歧义。不能把手工提供的正确结构当成模型已经学会的能力。

The experiment uses seven authored records and describes no real vendor or food service. D1 states that stall A's kung pao chicken contains chicken and peanuts. D2 assigns tofu and beef to stall B's mapo tofu. D3 reports long evening waits at A, and D4 high evening order volume there. D5 assigns chicken and lettuce to stall C's salad, D6 describes mild taste, and D7 explicitly excludes peanuts from A's kung pao chicken. To isolate conflict, D1 and D7 are assumed to share one effective version; a real system must verify that condition.

Relations are manually supplied. The program performs deterministic filtering, status classification, and arithmetic without language model extraction, visual recognition, or a GraphRAG framework. Results verify implementation of the supplied logic, not automatic construction, semantic retrieval, or generation quality. The small collection permits record by record inspection and classroom calculation. Future model integration should treat manual relations as reference annotations and separately test extraction errors,

ambiguous names, and versions. Correct structure provided by hand is not evidence that a model learned to construct it.

关系查询与来源回溯

Relation Queries and Source Traceback

查询“哪些记录中的菜品含鸡肉”时，程序筛选正向食材观察，得到窗口 A 宫保鸡丁和窗口 C 鸡肉沙拉，分别由 D1 和 D5 支持。这里回答的是资料中明确出现的关系，不是对全部菜品配方的常识推测。若把查询改为“哪些菜实际一定含鸡肉”，实验资料不足以支持更强措辞，因为这些记录只是人为设定的陈述。这个例子说明检索结果与现实保证之间还隔着来源可信度和适用范围。

回溯路径可以写成问题到食材、食材到观察、观察到菜品版本、观察到来源。每条答案都应带来源，避免把两个对象连接后丢失最初支持。假如删除 D5，窗口 C 沙拉的鸡肉关系应失去唯一支持并退出当前回答，而不是因为菜名中有鸡肉就自动继续保留。名称可以帮助召回候选，不能替代证据关系。若存在另一条独立确认，则应保留关系但更新支持集合。这个简单删除思想实验也解释了工程章依赖图的作用：系统需要知道结论依赖哪些观察，才能在来源变化后正确调整，而不只是保存一个固定答案。

The query asking which recorded dishes contain chicken filters positive ingredient observations and returns A's kung pao chicken and C's salad, supported by D1 and D5. It answers explicit relations in the records rather than inferring recipes from general knowledge. A stronger question asking which dishes certainly contain chicken in reality is not supported by authored statements alone. The example separates retrieval results from real world guarantees, which additionally require source reliability and applicability.

Traceback can connect the question to an ingredient, the ingredient to observations, observations to dish versions, and observations to sources. Each answer needs provenance. If D5 were deleted, the salad's chicken relation would lose its sole support and should leave the current answer rather than survive because chicken appears in its name. Names can retrieve candidates but cannot replace evidence. Another independent confirmation could preserve the relation with a changed support set. This deletion thought experiment illustrates dependency graphs: conclusions must retain their supporting observations so source changes can update them correctly instead of leaving a fixed answer.

冲突未知与因果边界

Conflict Unknowns and Causal Boundaries

程序对窗口 A 花生属性返回冲突，因为 D1 明确肯定而 D7 明确否定。它不会按多数、最新文件名或更长句子自动选一方。对窗口 B 花生属性返回未知，因为没有相关观察。未知不同于明确不含，这个区别是开放世界资料中的基本要求。如果后来获得版本信息，发现 D1 与 D7 适用于不同日期，原先冲突可被重新解释为时间变化。由此可见，冲突状态不是永久标签，而是相对于当前身份与时间假设的判断。

D3 与 D4 共同描述晚高峰现象，但实验没有建立因果边。程序可以把两条记录同时检索出来，答案应表述为资料同时提到等待久与订单多，不能写成订单多已经被证明导致等待久。要研究原因，

需要更多时间、流程和比较证据。这个例子也可用于检查生成模型是否擅自补充关系：给模型同样资料，观察其是否使用导致、因此或必然等超出支持的词。若实际运行这样的测试，必须保存提示、模型和输出版本。本报告只执行了确定性状态检查，没有把这一后续生成实验伪装成已完成结果。

The program returns conflict for peanuts in A's dish because D1 affirms presence and D7 explicitly denies it. It does not select a side by majority, filename recency, or sentence length. Peanuts in B's dish remain unknown because no relevant observation exists. Unknown is distinct from explicit absence. If version evidence later assigns D1 and D7 to different dates, apparent conflict can become temporal change. Conflict is therefore relative to current identity and time assumptions rather than a permanent label.

D3 and D4 describe evening conditions without establishing causation. Both can be retrieved, but the answer should report their coexistence rather than claim proof that order volume caused waiting. Causal investigation needs additional temporal, process, and comparison evidence. A future generation test could inspect whether a model adds words such as therefore or causes. Such a test would require saved prompts, model versions, and outputs. This report executes deterministic state checks only and does not present that proposed generation experiment as completed.

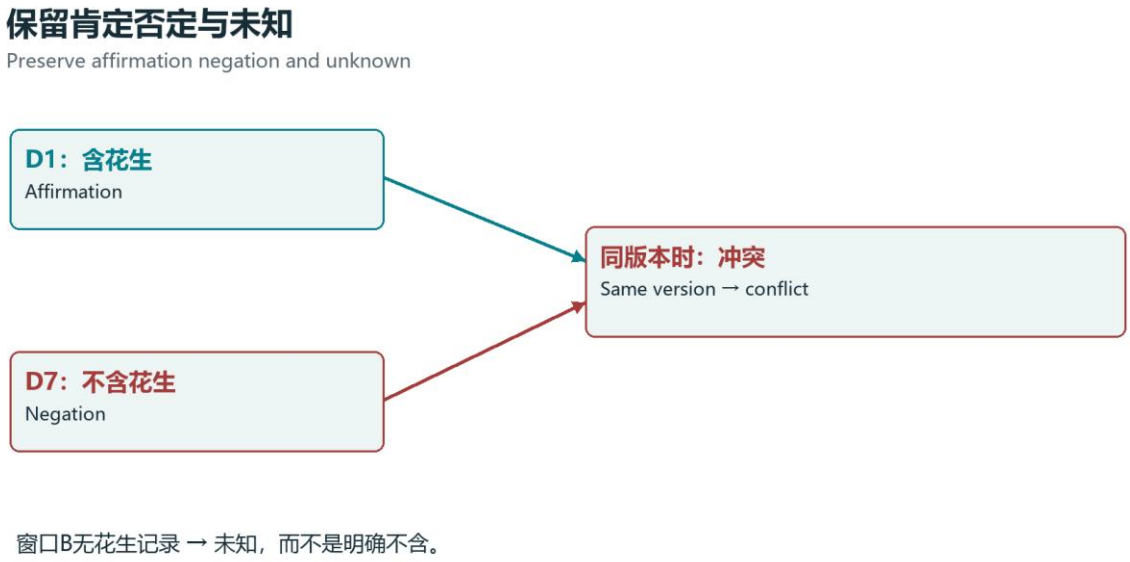


图 22 人工模拟资料中的冲突与未知 / Conflict and unknown in synthetic records
来源 / Source 本报告模拟 D1—D7；无模型推理 / Synthetic records, no model inference

排名指标的手工复算

Recalculate Ranking Metrics by Hand

排名演示返回五项，相关等级依次为 3、0、2、0、1；还有一项等级为 1 的相关候选未召回，完整集合因此有六项、四项相关。前五项中三个相关，精确率为 0.60，召回率为 0.75。按本章声明的指数增益与对数折扣，DCG 约为 8.886853，完整集合的理想前五项 3、2、1、1、0 的 IDCG 约

为 9.823466，nDCG 约为 0.904656。这些值由脚本计算，并与 Wolfram 计算结果核对。数据是为解释指标构造的，不是任何检索器的实测排名。

这个例子可说明精确率与 nDCG 关注不同方面。前五项包含两个不相关候选，使精确率不高；最高等级证据排在第一位，使分级排序分数仍较高。若把等级 3 移到最后，相关项总数不变，精确率与召回率不变，但 nDCG 会下降。若不知道全部相关项数量，就不能得到同样意义的召回率。实验记录因此保存相关等级、理想排序、相关总数和公式定义，而不仅保存一个最终分数。读者可以修改序列观察变化。指标解释必须与标注规则一起交付，否则相同数字可能对应完全不同的检索质量。

The ranking demonstration returns grades 3, 0, 2, 0, 1. Another relevant candidate with grade 1 is not retrieved, giving six candidates and four relevant items overall. Three relevant results among five give precision 0.60 and recall 0.75. Under the stated exponential gain and logarithmic discount, DCG is approximately 8.886853. The full collection's ideal top five, 3, 2, 1, 1, 0, gives IDCG approximately 9.823466 and nDCG approximately 0.904656. Script results were checked against Wolfram. These are constructed metric examples, not measured rankings from a retriever.

Precision and nDCG emphasize different properties. Two irrelevant candidates lower precision, while placing the most important evidence first keeps graded ranking quality relatively high. Moving grade 3 to the end preserves precision and recall but lowers nDCG. Without the total number of relevant items, the stated recall cannot be computed. The record therefore includes grades, ideal order, relevant total, and formula definitions rather than only a final score. Readers can modify the sequence and inspect changes. Metrics must be delivered with annotation rules because identical numbers can describe different retrieval conditions.

假设成本情景与盈亏平衡

Assumed Cost Scenarios and Break Even

成本演示设基线构建成本 100 单位、每次查询 1 单位，图方案构建 400 单位、每次查询 0.6 单位，暂忽略更新与运维。查询量为 100、500、1,000 和 5,000 时，两方案总成本分别为 200 与 460、600 与 700、1,100 与 1,000、5,100 与 3,400 单位。两条直线在 750 次查询处相等。这些参数是明确假设，不代表真实模型价格、实际性能或图方法普遍更省。计算只说明较高固定成本与较低边际成本如何随用量变化。

真实决策还需加入更新频率、图维护、人力审核和任务分布。如果图方案的单次查询并不更便宜，就可能不存在正向盈亏平衡点；如果其价值主要是减少严重错误，也不能只按计算费用比较。可以对构建成本与单次成本分别做敏感性分析，找出结论最依赖哪些假设。报告附带 CSV 与 JSON，读者可以替换参数重新计算。图中横轴必须写查询次数，纵轴写假设成本单位，并在标题中标明情景性质。成本可视化的作用是展示条件关系，而不是用一条交叉曲线证明某种架构必然值得采用。

The cost scenario assumes a baseline build of 100 units and one unit per query, versus a graph build of 400 and 0.6 per query, omitting updates and operations. At 100, 500, 1,000, and 5,000 queries, totals are respectively 200 versus 460, 600 versus 700, 1,100 versus 1,000, and 5,100 versus 3,400. Equality occurs

at 750 queries. These are assumptions, not prices, measured performance, or evidence that graphs generally cost less. The calculation illustrates fixed and marginal cost interaction.

Real decisions need update frequency, graph maintenance, review effort, and task distributions. If graph queries are not cheaper, no positive break even may exist. If their value is reduced serious error, compute fees alone are inadequate. Sensitivity analysis can vary construction and query costs separately to identify decisive assumptions. Attached CSV and JSON permit replacement values. Figures must label query counts and assumed cost units and state their scenario status. The visualization explains conditional relationships rather than proving that an architecture is necessarily worthwhile.

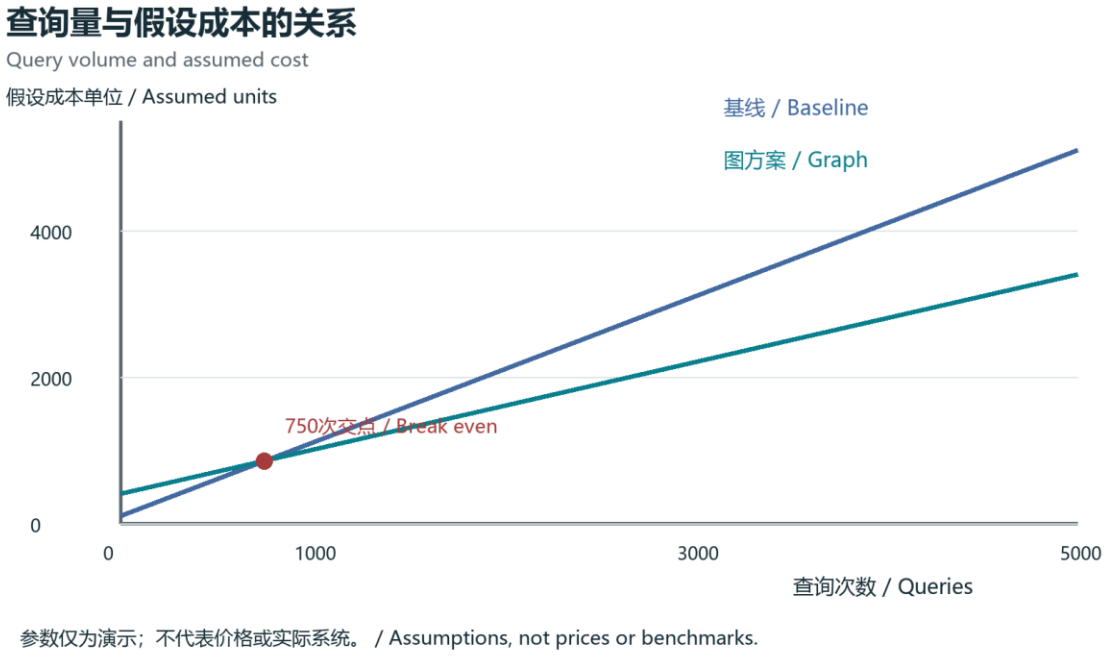


图 23 固定成本与边际成本的假设情景 / Assumed fixed and marginal costs
来源 / Source 本报告假设；数据见 CSV 与 JSON / Assumed inputs in attached data

从微型逻辑实验扩展到真实研究

Extend a Logic Demonstration into a Real Study

微型实验通过五项代码检查，覆盖鸡肉关系结果、花生冲突、另一对象的未知状态、冲突来源保留以及排名数值一致性。它的优点是可逐条解释，局限是没有自然语言理解、视觉噪声和大规模检索。扩展研究时，第一步应加入真实但可审核的小资料集，保持人工参考标注；第二步接入抽取与链接模型；第三步加入检索与生成；最后才扩大规模。每一步都保留上一阶段检查，避免最终回答掩盖中间错误。

建议在真实实验中增加同名不同对象、同对象不同日期、否定表达、缺失图片和错误配对。比较基线与图方法时固定资料、模型和预算，并记录额外构建成本。对于输出差异，逐条定位到候选、路径与引用，而不仅统计总分。若结果没有提升，可以检查图质量不足、路由错误或问题本身不需要关系结构等解释，再设计下一轮实验。研究价值来自可证伪的判断与可复查过程。一个小而

诚实的实验能够成为后续工作的可靠起点；一个未经执行却写满性能数字的实验则无法提供这种基础。

Five code checks cover the chicken relation result, peanut conflict, another object's unknown state, preservation of conflicting sources, and ranking consistency. The demonstration is individually explainable but contains no language understanding, visual noise, or large scale retrieval. Extension should first add a small auditable real collection with manual references, then extraction and linking models, then retrieval and generation, and only afterward scale. Earlier checks should remain so final answers do not conceal intermediate errors.

Real experiments should add same named distinct objects, one object across dates, negation, missing images, and wrong pairings. Baseline and graph comparisons should fix sources, models, and budgets and record construction overhead. Output differences should trace to candidates, paths, and citations rather than only total scores. No improvement can motivate tests of graph quality, routing, or whether the questions need structure. Research value comes from falsifiable judgments and inspectable processes. A small honest experiment can support later work; unexecuted performance numbers cannot.

返回目录 Back to contents

第16章　研究议程采用路线与综合判断

Research Agenda Adoption Roadmap and Synthesis

检验图结构究竟贡献了什么

Test What Graph Structure Actually Contributes

未来研究首先需要把图结构的贡献与其他改进分开。一个可检验假设是：在需要跨片段连接的问题上，具有正确类型和身份的关系，比仅提供相同实体名更有帮助。实验可以固定原文、生成模型和上下文预算，比较真实关系、随机关系、无关系和人工参考关系。如果真实图与无图差异很小，而人工参考图明显更好，瓶颈可能在构图质量；如果两种图都无改善，则可能是问题不需要图或生成器没有利用结构。

这种设计比单纯增加一个图模块更能解释结果。还可以按路径长度、实体歧义和关系稀有度分组，判断结构在哪些条件下有用。随机边对照应控制图密度，避免因为上下文长度不同引入混杂；人工图则需要记录标注成本，不能作为免费资源比较。研究报告应同时给出成功与失败问题的证据路径，让读者判断收益是否来自合理连接。图的价值不必表现为所有任务都提升。能够准确识别需要图的子任务，并在其他任务中回退到更简单方法，本身就是具有实用意义的研究结果。

Future work should isolate graph structure from accompanying improvements. One testable hypothesis is that typed relations with correct identity help cross passage questions more than the same entity names alone. Fix sources, generator, and context budget, then compare real relations, random relations, no relations, and manually annotated reference relations. If the real graph adds little while the reference graph helps, construction may be the bottleneck. If neither helps, the questions may not need structure or the generator may not use it.

This design explains more than simply adding a graph module. Grouping by path length, ambiguity, and relation rarity can identify useful conditions. Random controls should preserve density to avoid context length confounding. Reference graphs need annotation cost rather than being treated as free. Successful and failed evidence paths should accompany scores so readers can inspect whether connections are reasonable. Graph value need not appear on every task. Identifying the subset that benefits and using simpler retrieval elsewhere is itself a practical research result.

把不确定性传播到摘要与答案

Propagate Uncertainty into Summaries and Answers

当前系统容易在层层摘要中丢失不确定性：原文的可能变成观察中的肯定，观察中的分歧又在社区报告中被合并。一个重要研究方向是让知识状态沿派生链传播，而不是只在最后一句添加模糊免责声明。可以为观察保存支持、反对与未知条件，在生成摘要时要求保留分歧，最后检查答案是否维持同样边界。这里需要区分模型概率、来源可靠性和证据完整性，它们不是可以直接相乘的同一种数值。

可检验假设是，显式状态与证据依赖能降低冲突问题上的过度确定，同时不过度增加拒答。实验应同时包含一致、冲突、时间变化和资料缺失，分别衡量正确判断与合理回答范围。还应检查压缩程度是否影响少数证据保留。研究难点在于如何给用户足够信息而不造成阅读负担，可以比较简短状态标签、并列来源和展开式证据三种呈现。良好的不确定性表达不是堆叠可能、也许等词，而是说明哪些结论被支持、哪些条件尚未确认，以及补充什么证据能改变判断。

Uncertainty can disappear through successive summaries: possible statements become affirmative observations, and disagreements disappear inside community reports. Research should propagate knowledge states through dependencies rather than append vague caution at the end. Observations can retain support, opposition, and unknown conditions; summaries can preserve disagreement; answers can be checked for equivalent boundaries. Model probability, source reliability, and evidence completeness are different quantities and should not be multiplied as if they shared one interpretation.

A testable hypothesis is that explicit states and dependencies reduce unwarranted certainty on conflicts without excessive abstention. Experiments need consistent, conflicting, temporally changing, and incomplete cases, measuring both correct judgments and justified answer scope. Compression should be tested for minority evidence retention. Presentation research can compare concise state labels, parallel sources, and expandable evidence to balance clarity with reading burden. Good uncertainty communication identifies supported conclusions, unresolved conditions, and evidence that could change judgment, rather than merely repeating words such as perhaps.

研究模态必要性与证据充分性

Study Modality Necessity and Evidence Sufficiency

多模态 GraphRAG 需要回答一个比能否处理图片更严格的问题：哪些问题必须依赖图片，哪些图片区域真正改变答案。可以把模态必要性作为问题标注的一部分，要求标注者说明文字是否足够、视觉证据补充了什么、是否存在替代来源。然后比较系统在完整输入、移除视觉和错误视觉条件

下的变化。若移除图像后答案不变，需要判断是文字已经足够，还是模型忽视了视觉；仅凭分数稳定无法区分这两种情况。

证据充分性也可以研究为最小支持集合问题。对于一个陈述，哪些观察共同足以支持，哪些只是背景，哪些是反例？图结构可以表达这些角色，生成器则应在预算内保留关键组合。实验可测量减少证据数量时的质量曲线，而不是只测试固定前若干项。对视频与复杂页面，还要考虑时间或空间关系的必要性。研究目标不是让所有模态永远参与，而是让系统选择足够且必要的证据。这样既可能降低成本，也能提高解释清晰度，并减少无关图片或冗余图路径给模型带来的干扰。

Multimodal GraphRAG needs a stricter question than whether it accepts images: which questions require them and which regions change answers? Modality necessity can become an annotation field explaining whether text suffices, what vision contributes, and whether alternative evidence exists. Compare complete input, removed vision, and incorrect vision. Stable answers without images may mean text is sufficient or the model ignored vision; scores alone cannot distinguish these explanations.

Evidence sufficiency can be studied through minimum support sets. Which observations jointly support a claim, which provide background, and which challenge it? Graphs can encode these roles while generation preserves crucial combinations under budgets. Experiments can measure quality as evidence is removed rather than testing only one fixed top count. Video and complex pages also require spatial and temporal necessity analysis. The goal is sufficient and necessary evidence, not permanent participation of every modality. This can reduce cost, improve explanations, and limit distraction from irrelevant images or redundant paths.

动态知识与长期评测

Dynamic Knowledge and Longitudinal Evaluation

多数静态评测难以说明系统运行数月后是否仍可靠。长期研究可以按时间输入新资料、修订和撤回，定期查询同一组历史问题与新问题，观察知识更新、遗忘和引用有效性。关键指标包括从变更到生效的延迟、失效证据继续被引用的比例、实体误合并累积以及修复后的复发。这个方向需要保存完整版本，不然无法判断系统当时可见的资料是什么，也就无法公平评估历史回答。

长期实验还应测量维护工作，而不只测量自动化部分。人工审核时间、规则更新和故障处理可能成为实际成本主体。可以比较全量重建、局部更新和延迟更新策略，在同一变化序列下观察质量与成本。数据漂移应与软件升级分开记录，避免把两者混在一次前后比较中。研究结果若只能在一次固定快照上成立，应明确其时间边界；若能跨多轮更新保持稳定，则需要展示稳定性的证据。动态知识研究将 GraphRAG 从一次索引与回答，推进到可持续的知识服务，这也是其应用价值能否长期成立的重要条件。

Static tests rarely establish reliability after months of operation. Longitudinal studies can introduce new records, revisions, and withdrawals over time, repeatedly querying historical and new questions. Measures include change to availability delay, continued citation of invalid evidence, accumulated false merges, and recurrence after repairs. Complete versions are essential to reconstruct what information was available when an answer was produced and evaluate it fairly.

Studies should also measure maintenance rather than only automation. Human review, rule changes, and incident work may dominate costs. Full rebuilds, local updates, and delayed updates can be compared under one change sequence. Data drift and software upgrades should be recorded separately instead of combined into one before and after comparison. Snapshot findings need temporal boundaries; claims of stability across updates need evidence across those updates. Dynamic evaluation moves GraphRAG from a one time indexing exercise toward a sustainable knowledge service, a central condition for durable application value.

分阶段采用路线

A Staged Adoption Roadmap

建议采用四个有退出条件的阶段。第一阶段梳理资料与问题，建立来源清单、权限和人工参考答案；第二阶段实现可靠的基础检索与引用，明确哪些错误仍未解决；第三阶段只为关系、全局或视觉必要问题增加相应结构，并进行同条件对照；第四阶段加入持续更新、纠错与运营监测。每一步的交付应是可检查的结果，而不只是完成一个模块。阶段之间不必按固定日历推进，应由证据与团队能力决定。

进入下一阶段前可以问三个问题：新增能力是否改善目标任务，新增失败模式是否能被发现和修复，维护成本是否有责任主体承担。若任一问题答案不清楚，就应缩小范围或继续验证。对高校课程，路线可以转化为从纸卡到表格再到代码的学习序列；对企业，路线可以转化为小范围试点与逐步推广。图谱规模、模型参数和图像数量都不能代替验收结果。报告建议保留每阶段的决策记录，包括选择、证据、未解决问题和回退方式，使技术路线能够随着新的实验结果调整。真正成熟的采用过程允许简化，而不把复杂性本身当作进步。

A roadmap can use four stages with exit conditions. First inventory sources and questions, permissions, and reference answers. Second establish reliable basic retrieval and citations and identify remaining errors. Third add structure only for relational, global, or visually necessary questions and compare under matched conditions. Fourth introduce updating, correction, and operational monitoring. Deliverables should be inspectable outcomes rather than completed modules. Progress should follow evidence and team capacity instead of a fixed calendar.

Before advancing, ask whether the addition improves target work, whether new failures can be detected and repaired, and whether maintenance has an accountable owner. Unclear answers justify narrower scope or further testing. Universities can translate the roadmap into paper, table, and code learning stages; enterprises into bounded pilots and expansion. Graph size, model parameters, and image counts do not replace acceptance results. Decision records should preserve choices, evidence, unresolved issues, and fallback routes so the approach can change with experiments. Mature adoption permits simplification rather than treating complexity itself as progress.

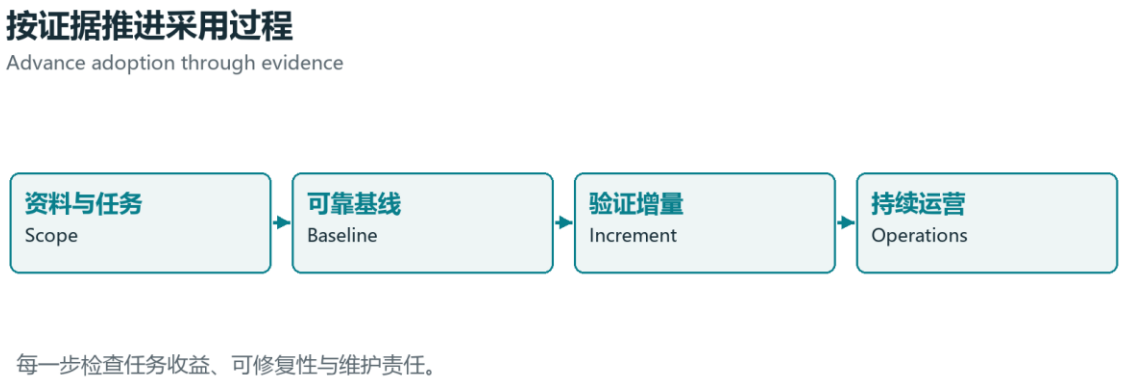


图 24 阶段间设置验收与退出条件 / Acceptance and exit conditions between stages
来源 / Source 本报告原创设计 / Original design

两条研究线的连接与最终判断

Connecting the Two Research Lines and the Final Assessment

微软文章与美团文章分别突出知识使用与知识构建的不同侧面。前者说明关系结构和社区层次可以帮助面向私有叙事资料的发现，后者说明食材级多模态信息可以支持细粒度菜品理解。把两者连接起来，需要补上统一身份、证据定位、版本、权限和评测，而不是只在系统名称前加上多模态。图谱为关系提供结构，多模态为观察提供互补，检索为问题选择材料，生成负责组织表达；任何一层的错误都可能改变最终判断。

本报告的综合判断是，GraphRAG 与多模态图谱值得作为一组可检验的工程与研究方法使用，其价值取决于问题确实需要关系或视觉证据、来源足够可靠、错误能够追踪以及维护能够持续。它们不会自动消除幻觉，也不会把相似度变成事实或把共现变成因果。最有说服力的应用不是展示最大网络，而是在具体问题上给出更完整的必要证据、更清楚的不确定性和更可修复的过程。后续研究应围绕这些可观察结果展开，并同时报告没有改善的条件。只有如此，技术的进步才能转化为更可靠的知识工作与更扎实的教学实践。

The Microsoft and Meituan articles emphasize different sides of knowledge use and construction. The former shows how relationships and community hierarchies can support discovery in private narrative material. The latter shows how ingredient level multimodal information supports detailed dish understanding. Connecting them requires shared identity, evidence locations, versions, permissions, and evaluation, not simply adding multimodal to a system name. Graphs structure relations, modalities provide complementary observations, retrieval selects material for questions, and generation organizes expression. Errors at any layer can change the final judgment.

The synthesis is that GraphRAG and multimodal graphs are useful as testable research and engineering methods when questions require relational or visual evidence, sources are adequate, errors are traceable, and maintenance is sustainable. They do not automatically eliminate hallucinations, turn similarity into fact, or convert cooccurrence into causation. Persuasive applications provide more complete necessary evidence, clearer uncertainty, and more repairable processes for concrete tasks rather than the largest

network. Future work should examine these observable outcomes and report conditions without improvement. That discipline connects technical progress to reliable knowledge work and substantive teaching practice.

返回目录 Back to contents

附录 A 研究与实施操作表

Appendix A Research and Implementation Worksheets

文献发现记录与筛选字段

Literature Discovery Logs and Screening Fields

文献发现记录应保存查询原文、检索日期、使用平台、筛选理由和最终主来源。可使用的问题组合包括 GraphRAG 与 evaluation、multimodal knowledge graph 与 entity linking、ingredient retrieval 与 CMIngre，以及与具体任务相关的词组。宽泛查询用于发现研究分支，精确标题查询用于定位原始论文与数据。记录被排除材料同样有价值，例如缺少方法描述、仅转载新闻或无法确认版本，能解释综述为何没有纳入某些结果。叙述性综述不需要伪装成穷尽检索，但应让选择过程可理解。

本报告使用 Consensus 和 SciSpace 发现论文线索，再回到论文页面、出版记录、作者数据集或官方文档核对。工具生成的摘要和限制描述不直接作为作者结论，尤其不能把对基线方法的批评误当成新方法自身的局限。书目中的读取范围区分完整文章、方法段和摘要记录，帮助解释论断的支持深度。进一步研究可以扩展数据库、设定预注册纳入规则，并由多人筛选，但那将是新的系统综述工作。操作表的重点是保留足够信息，让后续研究者知道应继续读什么、哪些结论还需要更深的原文核验，而不是把检索结果数量当成研究质量。

A discovery log should retain queries, dates, platforms, screening reasons, and final primary sources. Query combinations can include GraphRAG and evaluation, multimodal knowledge graph and entity linking, ingredient retrieval and CMIngre, plus task terms. Broad searches discover branches; exact titles locate originals and datasets. Exclusion reasons such as absent methods, news republication, or uncertain versions explain selection. A narrative review need not claim exhaustive retrieval, but its choices should be understandable.

This report uses Consensus and SciSpace for discovery and checks paper pages, publication records, author datasets, or official documentation. Generated summaries and limitations are not automatically author conclusions. Criticism of baselines must not become limitations attributed to a new method. Reading scope distinguishes articles, workflow sections, and abstract records. A future systematic review could add databases, preregister criteria, and use multiple screeners, but that is a separate study. The worksheet preserves what later researchers need to read and verify rather than treating result counts as quality.

陈述证据登记表

A Claim and Evidence Register

每条重要研究陈述可以登记为一行，字段包括陈述标识、具体文字、所属类型、来源、位置、适用条件、支持强度和核验状态。所属类型可分原文结果、跨文献综合、设计建议和演示计算。原文结果要保留实验范围，综合判断要列出推理依据，设计建议要说明待验证目标，演示计算要提

供输入与公式。这样同一段文字中的不同成分不会被一个笼统引用混在一起。登记表还可记录替代解释，帮助研究者主动寻找反例。

核验可以分两轮。第一轮检查来源是否存在、标题年份是否匹配、数字是否抄录正确；第二轮检查来源是否真正支持陈述，以及条件是否被扩大。对百分比，记录分母与单位；对比较结论，记录比较对象和配置；对不存在的证据，不用流畅措辞填空。若无法完成第二轮，应缩小陈述或保留为假设。双语报告还可以在同一行加入对应英文陈述，检查量词、否定与时间是否一致。登记表不是为了增加文书负担，而是降低长报告中条件丢失与引用漂移的风险，尤其适合多次修订和多人阅读的研究项目。

An important claim can occupy one register row containing identifier, wording, type, source, location, conditions, support strength, and verification status. Types include reported result, cross source synthesis, design recommendation, and demonstration calculation. Results retain experimental scope; syntheses list reasoning; recommendations state validation goals; calculations supply inputs and formulas. Distinct components in one paragraph then do not disappear behind one broad citation. Alternative explanations encourage active counterexample search.

Verification can use two passes. First check source existence, title and date, and copied numbers. Then check actual support and whether scope expanded. Percentages need denominators and units; comparisons need configurations. Missing evidence should not be filled with fluent prose. Incomplete support requires narrower wording or a hypothesis label. Bilingual claims can share a row to compare quantifiers, negations, and time. The register reduces lost conditions and citation drift in long and repeatedly revised reports rather than merely adding paperwork.

图谱数据契约检查表

A Graph Data Contract Checklist

数据契约应先说明对象类型与标识规则，再说明关系方向、限定条件和允许状态。以菜品版本为例，需要定义版本如何产生、何时生效、何时失效、是否允许重叠以及谁负责确认。食材观察需要区分肯定、否定与未知，不能只存布尔值后把缺失默认为假。资源定位要足够稳定，文件名改变不应导致全部引用失效。模型输出字段应包含方法版本，人工修改则应形成新记录，避免覆盖后无法追查。

接入测试可用少量故意设计的异常记录：重复标识、同名不同对象、缺失单位、时间逆序、无来源关系和互相矛盾的观察。系统应拒绝非法结构或将其隔离，不能因为成功写入数据库就判定质量合格。检查还应覆盖导出与重建，确认核心证据不依赖某个界面私有状态。对每个字段写清由谁提供、谁验证、下游如何使用和缺失时怎么办。契约变更需要版本与迁移说明。这样的表格使本体从抽象概念变成可执行约束，也让数据提供者与应用开发者拥有共同的质量语言。

A data contract defines object types and identity rules before relation direction, qualifiers, and allowed states. Dish versions need creation, effective and expiry rules, overlap policy, and confirmation responsibility. Ingredient observations distinguish affirmation, negation, and unknown rather than treating missing Boolean values as false. Resource locations should survive filename changes. Model outputs retain method versions, and human edits create new records instead of overwriting history.

Ingestion tests can include duplicate identifiers, same named distinct objects, missing units, reversed times, unsourced relations, and contradictory observations. Invalid structures should be rejected or isolated; a successful database write is not quality acceptance. Export and reconstruction should confirm evidence does not depend on private interface state. Every field needs an owner, validation rule, downstream purpose, and missing value behavior. Contract changes need versions and migration notes. The worksheet turns an ontology into executable constraints and gives data providers and developers a shared quality language.

实验运行卡与配置记录

Experiment Run Cards and Configuration Records

一次实验运行卡应包含目标假设、输入清单、划分规则、代码版本、模型标识、提示版本、随机种子、检索预算与执行时间。并非所有服务都能提供完全可重复结果，因此还要记录不可控因素，例如外部模型版本变化或近似搜索随机性。结果文件应与运行卡关联，失败运行也应保留状态。比较两次实验之前，先检查差异是否只有计划改变的因素，否则分数变化可能来自无意配置漂移。

运行卡还应区分已执行、部分执行和计划执行。某个模型调用失败后，不应使用手工猜测结果填补；某项消融尚未运行，就保持为空并说明待做。输出分析可以记录总体指标、分组结果、典型错误和成本，但结论要与实际执行范围一致。对本报告微型实验，运行卡明确人工关系、七条资料、确定性查询和五项代码检查；没有语言模型或视觉推理结果。读者可用附带脚本重新得到数值。真实系统研究则应增加依赖锁定和服务响应归档，在允许的隐私范围内保存必要输出。可复现不是承诺所有未来环境产生同一文本，而是提供足够条件解释结果并重新检验主要结论。

A run card records hypotheses, inputs, splits, code, model identifiers, prompts, seeds, retrieval budgets, and execution time. Services may not be fully reproducible, so uncontrollable factors such as external model changes and approximate search randomness also matter. Results link to run cards, including failed runs. Before comparison, confirm that only intended factors changed; otherwise configuration drift may explain score differences.

Cards should distinguish executed, partially executed, and planned work. Failed calls should not receive guessed outputs, and unrun ablations remain unreported. Analyses may include overall and grouped metrics, errors, and costs within actual execution scope. This report's run card specifies manual relations, seven records, deterministic queries, and five checks, with no language or visual inference results. Scripts reproduce the numbers. Real studies can add dependency locks and permitted response archives. Reproducibility means providing conditions sufficient to explain and retest main conclusions, not promising identical text in every future environment.

图表数据与视觉核验表

Figure Data and Visual Verification Worksheets

每张研究图应登记目的、数据来源、单位、计算步骤、图形类型和读者需要看到的结论。关系图要给边标签和方向，数量图要给坐标与基数，流程图要区分顺序与数据依赖。颜色应保持稳定含

义，但重要区分还需文字或线型，确保黑白打印可读。来源图与原创重绘应分别说明，局部裁剪保留必要上下文。若图使用假设数据，标题或图注必须直接标出，而不把说明藏在远处正文。

核验分内容与版面两层。内容检查数字、单位、标签和正文引用是否一致，是否把相关画成因果，是否遗漏例外；版面检查字号、裁切、重叠、清晰度和分页。图不应只在屏幕放大后才可读，嵌入正文后的实际尺寸才是验收尺寸。复杂图可以拆成总览与局部，避免把整套系统塞进一页。对于可编辑源文件，应保留文字和形状信息，而不仅提供低分辨率截图。报告附带 SVG 图源与数据，使后续备课可以修改标签和重组图示。视觉质量的标准是帮助读者准确判断，而不是装饰数量或色彩丰富程度。

Each figure needs a purpose, source, units, calculation steps, chart type, and intended observation. Relationship diagrams need labeled directions, quantity plots axes and bases, and workflows a distinction between sequence and dependency. Stable colours should be reinforced by words or line types for monochrome reading. Source figures and original redrawings need separate attribution, and crops must retain context. Assumed data should be labeled directly in titles or captions rather than only in distant prose.

Verification has content and layout layers. Content checks values, units, labels, textual references, causal implications, and exceptions. Layout checks type size, cropping, overlap, resolution, and pagination at actual embedded size. Complex figures can split into overview and detail rather than compress an entire system onto one page. Editable sources should retain text and shapes instead of only low resolution screenshots. Attached SVG and data support later teaching adaptation. Visual quality means helping accurate judgment rather than maximizing decoration or colour.

试点决策与复审记录

Pilot Decisions and Review Records

试点决策表可在同一页记录目标任务、当前基线、预期增益、关键风险、资源需求和停止条件。预期增益应写成可观察结果，例如减少遗漏证据或缩短核验时间，而不是泛称提升智能化。关键风险应来自场景中的实际失败，不必堆叠与项目无关的抽象清单。停止条件要能触发行动，例如某类关键错误持续存在时缩小回答范围，数据维护无人负责时暂不扩展。这样的决策表让技术与管理讨论围绕同一组事实。

复审时比较原先假设与当前证据，记录保留、调整或停止的理由。一次未达标不必意味着技术永远无用，但也不应通过不断更换目标使试点永远成功。可以预先规定主要指标与可接受变化，再把探索性发现单独报告。若决定采用，记录维护责任和下一次复审时间；若决定简化，保留可复用的资料清单、证据标识与测试集。研究成果不只包括最终系统，也包括对哪些条件有效的清楚认识。透明决策有助于团队积累经验，并使后来者能够在新的数据和模型条件下重新评估，而不是重复一次无法解释的试验。

A pilot decision sheet can record target tasks, baseline, expected gains, critical risks, resources, and stopping conditions. Gains should be observable, such as fewer evidence omissions or shorter verification time, rather than generic intelligence claims. Risks should follow actual scenario failures. Stopping

conditions need actions, such as narrowing scope after persistent critical errors or delaying expansion without maintenance ownership. This gives technical and management discussion a common factual basis.

Reviews compare initial assumptions with current evidence and record reasons to retain, adjust, or stop. Failure does not prove permanent uselessness, but changing goals repeatedly should not make every pilot successful. Define primary measures and acceptable changes beforehand and report exploratory findings separately. Adoption needs maintenance ownership and a next review date; simplification can preserve inventories, identifiers, and tests. Research outcomes include knowing which conditions work, not only a final system. Transparent decisions accumulate experience and permit reassessment under new data and models instead of repeating an unexplained trial.

返回目录 Back to contents

附录 B 双语术语与易混概念

Appendix B Bilingual Terms and Common Confusions

知识图谱实体与关系

Knowledge Graph Entity and Relation

知识图谱即 knowledge graph，是以明确对象及其关系组织信息的一类表示方式。实体即 entity，表示在当前建模范围内可以区分的对象；关系即 relation，表示对象之间有语义的联系。节点 node 是图结构中的位置，并不必然等于现实实体，它也可以表示文档、观察或事件。边 edge 是结构连接，只有定义类型与方向后才知道其含义。理解这些层次，可以避免把图上画出的所有连接都当成已经确认的现实关系。

例如，图片节点连接到菜品节点，可能表示展示该菜、疑似拍到该菜或由该菜商品页面提供，三种关系不能共用一个含糊的相关标签。本体 ontology 规定允许表达哪些对象与关系，模式 schema 规定具体数据结构，两者在工程中可能重叠，但不应把字段表等同于全部领域知识。术语使用应服务清晰表达：讨论对象身份时说实体，讨论算法结构时说节点，讨论证据时说观察或陈述。这样的区分也有助于跨语言阅读，因为英文论文有时在不同上下文中使用同一个 graph 词汇，实际研究对象却可能是实体网络、片段网络或文档网络。

A knowledge graph organizes information through explicit objects and semantic relations. An entity is a distinguishable object within a modeling scope. A relation expresses a meaningful connection. A node is a structural position and may represent an entity, document, observation, or event. An edge is a connection whose meaning requires type and direction. These distinctions prevent every drawn connection from being treated as a confirmed real world relationship.

An image linked to a dish may depict it, possibly depict it, or merely come from its product page. Those meanings should not collapse into related to. An ontology defines expressible concepts and relations, while a schema defines concrete structure; they overlap but a field list is not all domain knowledge. Use entity for identity, node for structure, and observation or claim for evidence. This also helps read papers whose graphs may actually be entity, chunk, or document networks.

检索增强与图增强的范围

The Scope of Retrieval Augmentation and Graph Augmentation

RAG 即 retrieval augmented generation，通常指利用检索到的外部材料支持生成。GraphRAG 在广义上表示图结构参与这一过程，但具体参与位置可能不同。检索 retrieval 负责找到候选，重排 reranking 负责重新排序，生成 generation 负责组织输出。三者可以由不同模型完成，也可以部分采用规则。检索到了正确材料不保证生成正确，生成流畅也不能证明检索充分。因此讨论系统效果时，要说明所指的是哪个阶段以及评价对象。

微软 GraphRAG 是一个具体项目与方法系列，不能用其名称覆盖所有图增强研究。全局搜索 global search 侧重跨较大范围的归纳，局部搜索 local search 侧重围绕相关对象组织材料；这里的全局不意味着系统掌握世界全部知识，局部也不意味着只读一份文件。多跳 multihop 表示回答需要连接多个证据步骤，不等同于回答很长或推理文字很多。一个只需要查表的问题，即使模型写出十段思考，也不会因此变成多跳任务。术语应根据证据结构使用，而不是根据输出外观或宣传措辞判断。

Retrieval augmented generation uses external retrieved material to support generation. GraphRAG broadly means that graph structure participates, but the point of participation varies. Retrieval finds candidates, reranking reorders them, and generation organizes output. Different models or rules may implement these stages. Correct evidence does not guarantee correct generation, and fluent generation does not establish sufficient retrieval. Effectiveness claims need an explicit stage and evaluation object.

Microsoft GraphRAG is a specific project and method family, not a name for every graph augmented study. Global search synthesizes a broader scope; local search organizes material around relevant objects. Global does not mean all world knowledge, and local does not mean one file. Multihop questions require connecting evidence steps. Lengthy reasoning prose does not turn a lookup into a multihop task. Terms should follow evidence structure rather than output appearance or promotional wording.

对齐链接融合与补全

Alignment Linking Fusion and Completion

对齐 alignment 表示不同表示之间建立对应，可能发生在图文嵌入、实体集合或局部区域层面。链接 linking 通常强调把提及或观察指向某个实体。融合 fusion 指结合多源信息，但融合结果可以保留冲突，不必只有一个值。补全 completion 指预测缺失关系或属性，这种预测是候选知识，不自动成为已确认事实。四个词都涉及连接，却回答不同问题：对应什么、指向谁、怎样结合、还可能缺少什么。

以菜品图片为例，文字与图像在共享空间相近属于表示对齐，把图片定位到窗口 A 商品属于实体链接，把菜单与图片观察共同登记属于证据融合，根据已知配方猜测另一种食材属于补全。若把最后一步的预测直接写回确认属性，就会把模型偏好变成循环证据。后续检索再次读到这个属性时，可能误以为存在独立来源。为避免这种自我强化，系统应保存产生方式与状态，训练标签、模型预测和人工确认也应分别管理。研究评价中，链接准确率、融合收益和补全排名不能互相替代；它们需要不同参考答案和错误分析。

Alignment establishes correspondence between representations, entity sets, or regions. Linking usually assigns a mention or observation to an entity. Fusion combines sources and may retain conflicts instead of producing one value. Completion predicts missing attributes or relations, creating candidates rather than confirmed facts. These terms ask different questions: what corresponds, which entity is meant, how sources combine, and what might be missing.

For a dish image, shared space similarity is alignment, assignment to stall A's product is linking, joint registration of menu and image observations is fusion, and guessing another ingredient from a recipe pattern is completion. Writing that guess into confirmed attributes creates circular evidence. Later

retrieval may mistake it for independent support. Systems should preserve production methods and states and distinguish training labels, predictions, and human confirmations. Linking accuracy, fusion benefits, and completion ranking require different reference answers and analyses.

依据忠实性与事实正确性

Grounding Faithfulness and Factual Correctness

Grounding 在本报告中指回答与可定位依据建立联系，faithfulness 指回答没有超出给定证据或歪曲其内容，factual correctness 则关心陈述与现实或可靠参考是否一致。三者相互相关但不等价。一段回答可以有出处，却引用了不支持它的段落；也可以忠实复述一份错误记录；还可能凭模型记忆说对事实，却没有提供当前任务要求的依据。评价时应分别检查支持关系、忠实程度和事实状态，而不把任何一种单独当成全部可靠性。

Hallucination 常被译为幻觉，但实际错误应进一步定位为无依据新增、错误归因、条件丢失或虚构来源。Abstention 表示在证据不足时不作确定回答，合理拒答不等于系统失败；过度拒答则可能阻碍本可完成的任务。Calibration 表示确信表达与实际正确程度相匹配，模型给出 0.9 并不意味着在该业务上已经校准。对用户最有帮助的表达往往不是一个精细分数，而是明确的支持来源与缺口。例如“已找到当前菜单记录，但图片版本未确认”比没有解释的高置信标签更可判断。这些术语共同服务于把回答变成可检查陈述。

Grounding connects an answer to locatable evidence. Faithfulness means not distorting or exceeding supplied material. Factual correctness concerns agreement with reality or reliable references. These are related but distinct. An answer can cite an irrelevant paragraph, faithfully repeat an incorrect record, or recall a correct fact without evidence required by the task. Evaluation should separately inspect support, faithfulness, and factual status.

Hallucination should be diagnosed more specifically as unsupported additions, misattribution, lost conditions, or invented sources. Abstention can be justified by insufficient evidence, while excessive abstention blocks answerable work. Calibration concerns agreement between expressed confidence and actual correctness; a score of 0.9 is not automatically calibrated for a business domain. Users often benefit more from explicit sources and gaps than a precise score. A current menu with an unconfirmed image version is more informative than an unexplained high confidence label. These concepts make answers inspectable.

精度召回覆盖与百分点

Precision Recall Coverage and Percentage Points

Precision 在检索中通常表示返回项中的相关比例，在目标检测中还涉及匹配规则与阈值；中文精度或准确率的译法需要结合原文定义。Recall 表示相关项被找回的比例，coverage 则可能表示对象、属性或任务的覆盖，不能仅因中文都涉及覆盖就视为同一指标。Accuracy 通常按明确评测单位计算正确比例，必须知道分母。Average precision 与 nDCG 是不同排序或检测指标，也不能只因为数值都在零到一之间就直接比较。

百分比变化需要区分百分点与相对变化。从 80%到 87.01%是增加 7.01 个百分点，相对增加约 8.76%，两种表述都可正确，但回答的问题不同。错误率由 20.38%降到 2.3%，是下降 18.08 个百分点，相对下降约 88.71%，不能写成下降 18.08%。报告涉及指标时应保留原文名称与定义，必要时加上公式与分母说明。图表纵轴也应明确是百分数、比例还是百分点差值。细小的单位错误可能大幅改变读者对效果的理解，因此数值核验不仅检查数字抄对，还要检查它在句子中的意义是否正确。

Precision usually means the relevant share of retrieved results, while detection also requires matching rules and thresholds. Translation depends on the original definition. Recall measures recovered relevant items; coverage may concern objects, attributes, or tasks and is not automatically recall. Accuracy requires a defined unit and denominator. Average precision and nDCG are different measures and cannot be compared merely because both may lie between zero and one.

Percentage points differ from relative change. Moving from 80% to 87.01% adds 7.01 points or approximately 8.76% relatively. Error falling from 20.38% to 2.3% drops 18.08 points or approximately 88.71% relatively. Reporting a drop of 18.08% would change the meaning. Preserve metric names and definitions, adding formulas and denominators when needed. Axes should distinguish percentages, proportions, and point differences. Numerical checking concerns both copied digits and their meaning in context.

复现泛化因果与应用价值

Reproduction Generalization Causality and Application Value

复现 reproduction 通常指依据已报告方法重新检验结果，但不同研究对重复性与可复现性的术语约定可能不同，因此应直接说明代码、数据和条件。泛化 generalization 指在未见条件下保持能力，需要说清新的是实体、来源、时间还是任务。消融 ablation 通过移除或改变组件研究贡献，公平消融需要控制其他条件。基准 benchmark 提供共同任务与规则，但一个基准上的高分不能独立证明真实部署价值。

相关 association 表示观察到联系，因果 causality 要求能够支持干预或原因判断的更强设计。图上有向边也不自动成为因果边；方向可能只表示陈述结构或时间次序。应用价值 utility 则取决于用户任务、错误后果、时间和成本，未必与通用分数完全一致。一个系统即使平均答题提升，也可能因版本错误而不适合服务当前业务；一个较简单系统可能在高频精确查询中更有价值。报告因此坚持把技术能力、测量结论和采用决策分开：能力需要实验，结论需要边界，采用需要任务与维护证据。清楚使用这些词，能够减少研究传播中的夸大，并让中英文读者围绕同一个问题讨论。

Reproduction retests reported findings, but terminology conventions vary, so state code, data, and conditions directly. Generalization requires specifying what is unseen: entities, sources, time, or tasks. Ablation studies component contributions while controlling other factors. Benchmarks provide shared tasks and rules, but high scores alone do not establish deployment value.

Association records a connection; causality requires stronger support for causes or interventions. Directed graph edges may express statement structure or temporal order rather than causation. Utility depends on tasks, consequences, time, and cost and may diverge from generic scores. Average gains can coexist with

disqualifying version errors, while simpler systems may better serve frequent precise queries. Technical capability, measured conclusions, and adoption are therefore distinct: capability needs experiments, conclusions need boundaries, and adoption needs task and maintenance evidence. Careful terminology reduces exaggerated communication and helps Chinese and English readers discuss the same question.

返回目录 Back to contents

参考文献

References

以下条目按引用编号排列。论文、预印本、官方文档、技术文章和作者数据集分别标明；出版与访问日期按可核对记录保留。 / Entries identify publication type and version dates.

[R01] Jonathan Larson and Steven Truitt. GraphRAG: Unlocking LLM discovery on narrative private data. 2024-02-13. Microsoft Research blog. 原始来源 / Primary source Accessed 2026-10-18.

[R02] Darren Edge et al.. From Local to Global: A GraphRAG Approach to Query-Focused Summarization. 2024; v2 2025-02-19. Research preprint. arXiv 2404.16130 Accessed 2026-10-18.

[R03] 美团技术团队. 基于多模态信息抽取的菜品知识图谱构建. 2024-05-17. Industry technical article. 原始来源 / Primary source Accessed 2026-10-18.

[R04] Lanjun Wang et al.. Toward Chinese Food Understanding: A Cross-Modal Ingredient-Level Benchmark. 2024 online; IEEE TMM volume 27 (2025), pp. 2863–2874. IEEE Transactions on Multimedia. DOI 10.1109/TMM.2024.3387735 Accessed 2026-10-18.

[R05] CMIngre research team. CMIngre Dataset and Supplementary Material. accessed 2026-10-18. Author dataset. 原始来源 / Primary source Accessed 2026-10-18.

[R06] Patrick Lewis et al.. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2020. Research paper. arXiv 2005.11401 Accessed 2026-10-18.

[R07] Boci Peng et al.. Graph Retrieval-Augmented Generation: A Survey. 2024 preprint; later journal version. Survey. arXiv 2408.08921 Accessed 2026-10-18.

[R08] Haoyu Han et al.. Retrieval-Augmented Generation with Graphs (GraphRAG). 2025 arXiv identifier. Survey. arXiv 2501.00309 Accessed 2026-10-18.

[R09] Haoyu Han et al.. RAG vs. GraphRAG: A Systematic Evaluation and Key Insights. 2025 preprint. Comparative study. arXiv 2502.11371 Accessed 2026-10-18.

[R10] Zirui Guo et al.. LightRAG: Simple and Fast Retrieval-Augmented Generation. 2024 preprint; 2025 EMNLP Findings. Research paper. arXiv 2410.05779 Accessed 2026-10-18.

[R11] Bernal Jiménez Gutiérrez et al.. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. 2024. Research paper. arXiv 2405.14831 Accessed 2026-10-18.

[R12] Bernal Jiménez Gutiérrez et al.. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. 2025. HippoRAG 2 research paper. arXiv 2502.14802 Accessed 2026-10-18.

[R13] Xiangrong Zhu et al.. Knowledge Graph-Guided Retrieval Augmented Generation. 2025. KG2RAG research paper. arXiv 2502.06864 Accessed 2026-10-18.

[R14] Xiaoxin He et al.. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. 2024. Research paper. arXiv 2402.07630 Accessed 2026-10-18.

[R15] Xiangru Zhu et al.. Multi-Modal Knowledge Graph Construction and Application: A Survey. 2022 preprint. Survey. arXiv 2202.05786 Accessed 2026-10-18.

[R16] Zhuo Chen et al.. Knowledge Graphs Meet Multi-Modal Learning: A Comprehensive Survey. 2024. Survey. arXiv 2402.05391 Accessed 2026-10-18.

[R17] Alec Radford et al.. Learning Transferable Visual Models From Natural Language Supervision. 2021. ICML research paper. 原始来源 / Primary source Accessed 2026-10-18.

[R18] Wenhu Chen et al.. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. 2022. EMNLP research paper. 原始来源 / Primary source Accessed 2026-10-18.

[R19] Xueyao Wan and Hang Yu. MMGraphRAG: Bridging Vision and Language with Interpretable Multimodal Knowledge Graphs. 2025 first version; v3 2026-07-18. Research preprint. arXiv 2507.20804v3 Accessed 2026-10-19.

[R20] Xu Yuan, Liangbo Ning, Qingqing Ye, Wenqi Fan and Qing Li. mKG-RAG: Leveraging Multimodal Knowledge Graphs in Retrieval-Augmented Generation for Knowledge-intensive VQA. 2025 preprint; v2 2026-04-27; SIGIR 2026. SIGIR 2026 paper; inspected arXiv version. arXiv 2508.05318v2 Accessed 2026-10-19.

[R21] Yin Wu, Quanyu Long, Jing Li, Jianfei Yu and Wenya Wang. Visual-RAG: Benchmarking Text-to-Image Retrieval Augmented Generation for Visual Knowledge Intensive Queries. 2025 preprint v2; related SIGIR 2026 Resource Track version. Benchmark paper; inspected arXiv v2 and author repository. arXiv 2502.16636v2 Accessed 2026-10-19.

[R22] Juyeon Kim, Geon Lee, Taeuk Kim and Kijung Shin. KGMEL: Knowledge Graph-Enhanced Multimodal Entity Linking. 2025. SIGIR 2025 short paper; inspected arXiv record. arXiv 2504.15135 Accessed 2026-10-19.

[R23] Manuel Faysse et al.. ColPali: Efficient Document Retrieval with Vision Language Models. 2024. Research paper. arXiv 2407.01449 Accessed 2026-10-18.

[R24] Zirui Guo et al.. RAG-Anything: All-in-One RAG Framework. 2025. Research preprint. arXiv 2510.12323 Accessed 2026-10-18.

[R25] Shahul Es et al.. RAGAs: Automated Evaluation of Retrieval Augmented Generation. 2024. EACL System Demonstrations. 原始来源 / Primary source Accessed 2026-10-18.

[R26] Hao Yu et al.. Evaluation of Retrieval-Augmented Generation: A Survey. 2024. Survey preprint. arXiv 2405.07437 Accessed 2026-10-18.

[R27] V. A. Traag, L. Waltman and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. 2019. Scientific Reports. 原始来源 / Primary source Accessed 2026-10-18.

[R28] W3C. PROV Overview. 2013. W3C Working Group Note. 原始来源 / Primary source Accessed 2026-10-18.

[R29] Nelson F. Liu et al.. Lost in the Middle: How Language Models Use Long Contexts. 2023 preprint. Research paper. arXiv 2307.03172 Accessed 2026-10-18.

[R30] Alon Talmor et al.. MultiModalQA: Complex Question Answering over Text, Tables and Images. 2021. Benchmark paper. arXiv 2104.06039 Accessed 2026-10-18.

[R31] Yingshan Chang et al.. WebQA: Multihop and Multimodal QA. 2021 preprint; CVPR 2022. Benchmark paper. arXiv 2109.00590 Accessed 2026-10-18.

[R32] Microsoft. GraphRAG Indexing Dataflow. accessed 2026-10-18. Official documentation. 原始来源 / Primary source Accessed 2026-10-18.

[R33] Microsoft. GraphRAG Global Search. accessed 2026-10-18. Official documentation. 原始来源 / Primary source Accessed 2026-10-18.