

417 A Additional experiments

418 A.1 Topic-specific reasoning analysis

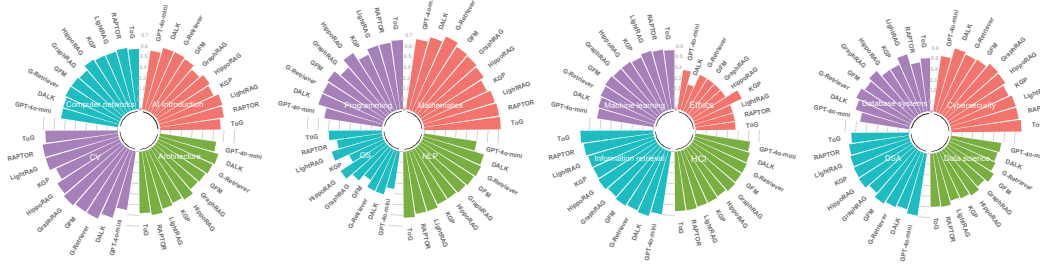


Figure 4: Comparison of R score by Topic.

419 Given that our dataset encompasses 16 thematic domains, we conducted experiments to analyze the
 420 reasoning capabilities of GraphRAG across different topics. Results indicate that the large language
 421 model (LLM) based on GPT-4o-mini demonstrates significant improvements in reasoning through
 422 GraphRAG across most domains. However, the following intriguing observations were made:

423 **Operating System Domain.** The LLM exhibits suboptimal performance in this domain. While
 424 GraphRAG provides marginal improvements in reasoning capabilities, overall scores remain low.
 425 This is primarily attributed to the highly specialized, systematic, and logically complex nature of
 426 operating system knowledge, which involves multi-layered principles such as process scheduling,
 427 memory management, and file systems, requiring precise grasp of conceptual definitions, algorithmic
 428 workflows, and causal relationships between entities. General-purpose training data for LLMs often
 429 lack comprehensive coverage of such granular knowledge systems, and the models themselves have
 430 inherent limitations in structured logical reasoning.

431 **Ethics Domain.** Consistent with the generation accuracy results, LLMs face substantial challenges
 432 in reasoning about ethical questions. Ethical problems fundamentally involve subjective value
 433 judgments, whose meanings are rooted in dynamic contexts of moral trade-offs and social norms. The
 434 symbolic representations captured by LLMs through statistical learning struggle to accurately model
 435 ambiguous ethical constructs, leading to intrinsic difficulties in both generating correct answers and
 436 constructing valid reasoning chains.



Figure 5: Comparison of AR score by Topic.

437 We further evaluated the AR scores of GraphRAG across different topics. Experimental results
 438 show that AR scores generally align with R scores in most cases. However, a notable observation
 439 emerges in the database systems domain: AR scores are significantly lower than R scores, indicating
 440 a high prevalence of "correct reasoning but incorrect answering" in LLMs, where reasoning steps
 441 diverge from final answer generation. This discrepancy arises because database system problems
 442 require models to reference specialized concepts such as relational algebra operations, transaction
 443 isolation levels, ACID properties, and query optimizer cost models, yet models do not perform
 444 formal computations or analyze critical factors like underlying data distribution and index selectivity.
 445 Although models may decompose processes like schema design or concurrency control according
 446 to human logical paradigms in chain-of-thought reasoning, their token selection during answer

447 generation prioritizes statistical fluency from training corpora over contextual logical accuracy. The
 448 strict requirements for precise logical operations (e.g., cost estimation, deadlock detection) in database
 449 tasks create a fundamental mismatch with the model’s learned fuzzy statistical patterns from general
 450 text, leading to reasoning chains that appear plausible in intermediate steps but produce erroneous
 451 conclusions at technical junctures, such as failing to execute physical query optimization calculations,
 452 due to the absence of real-world logical validation.

453 **A.2 Compute resources**

454 All code is done in Python, and experiments are conducted on H100*2 GPUs.

455 **B Metrics details**

456 **Prompt of OE and FB questions.**

Fig. 6 is the prompt used to generate the LLM-judge score for OE and FB questions.

You are a strict evaluator. Compare the following two answers for correctness and completeness:

Predicted Answer: {pred_answer}

Gold Answer: {gold_answer}

Please evaluate the predicted answer in comparison to the gold answer. Respond with a score between 0 and 1:

- 1: The predicted answer fully aligns with the gold answer.
- 0.5: The predicted answer is partially correct but lacks completeness or includes incorrect information.
- 0: The predicted answer is incorrect or completely misaligned with the gold answer.

Figure 6: Prompt of generation grading for OE and FB questions.

457

458 **Prompt of reasoning grading.**

459 Fig. 7 is the prompt used to evaluate the reasoning score R.

You are a strict evaluator. Compare the following two rationales for correctness and completeness:

Predicted Rationale: {pred_rationale}

Gold Rationale: {gold_rationale}

Please evaluate the predicted rationale in comparison to the gold rationale. Respond with a score between 0 and 1:

- 1: The predicted rationale fully aligns with the gold rationale.
- 0.5: The predicted rationale is partially correct but lacks completeness or includes incorrect information.
- 0: The predicted rationale is incorrect or completely misaligned with the gold rationale.

Figure 7: Prompt of rationale grading.

460 **Details of AR score.**

461 The AR score is computed based on the combination of answer correctness (generation score) and
 462 rationale correctness (reasoning score), with the following evaluation rules:

- 463 • When both the answer and rationale are fully correct (generation score = 1 and reasoning
 464 score = 1), the AR score is 1.0.
- 465 • If the answer is correct but the rationale is partially correct (generation score = 1 and
 466 reasoning score = 0.5), the AR score is 0.5.
- 467 • When the answer is correct but the rationale is incorrect (generation score = 1 and reasoning
 468 score = 0), the AR score is 0.0.

- 469 • For incorrect answers with a fully correct rationale (generation score = 0 and reasoning
470 score = 1), the AR score is 0.5.
- 471 • If both the answer is incorrect and the rationale is partially correct (generation score = 0 and
472 reasoning score = 0.5), the AR score is 0.25.
- 473 • In all other cases (e.g., incorrect answer with incorrect or missing rationale), the AR score is
474 0.0.

475 This scoring scheme systematically captures the alignment between answers and their supporting
476 reasoning, emphasizing the importance of both correctness and logical consistency in evaluating
477 model performance.

478 C Limitations of Existing GraphRAG Datasets

479 Through a systematic review of benchmark datasets used by contemporary Graph-RAG methods,
480 we have identified four critical limitations that undermine both task suitability and the validity of
481 evaluation results:

482 **Superficial retrieval tasks.** Most datasets pose questions that can be answered by straightforward text
483 retrieval, without requiring deep integration of graph structure or sophisticated semantic reasoning.
484 Consequently, models may achieve high scores by exploiting shallow keyword matching, offering no
485 insight into their true capabilities in relational reasoning or entity-association modeling.

486 **Synthetic and unrepresentative queries.** Questions are typically generated via hand-crafted rules,
487 yielding simplified language that lacks the domain-specific terminology, ambiguous intent, and
488 syntactic variety found in real user queries. This synthetic distribution diverges sharply from natural
489 problem settings, limiting the ecological validity of any conclusions about model generalization.

490 **Cross-task misalignment.** Many datasets are inherited from disparate tasks (e.g., knowledge-graph
491 question answering) whose annotation schemes and answer formats do not align with the core
492 objectives of Graph-RAG—namely, constructing and leveraging heterogeneous graph structures
493 to guide multi-source information fusion. Transferring evaluation metrics across tasks therefore
494 introduces inconsistencies that dilute the relevance of experimental findings for advancing Graph-
495 RAG techniques.

496 **Opaque reasoning evaluation.** Existing benchmarks supply only final answers or explicit node
497 sequences, but omit any structural or narrative annotation of the underlying inference process. Key
498 decision points—such as why a particular graph subpath was selected or how evidence from multiple
499 sources is reconciled—remain unexamined. Without annotated rationales, evaluation reduces to
500 binary correctness checks and cannot assess a model’s genuine reasoning competence.

501 These limitations collectively motivate the design of a dedicated benchmark that both challenges
502 Graph-RAG models on core reasoning skills and provides richly structured annotations for fine-
503 grained, interpretability-driven evaluation.

504 D Limitations of this paper

505 Despite the valuable contributions of this study, we acknowledge its limitations: (1) Our dataset
506 currently only contains English content; more detailed research should be done in the future for
507 different languages. (2) Other modal data such as images are not included in the current data set, and
508 richer multimodal datasets can be considered in the future.