# Characterizing Knowledge Graph Tasks in LLM Benchmarks Using Cognitive Complexity Frameworks

Sara Todorovikj[1,*], Lars-Peter Meyer[1,2] and Michael Martin[1,2]

[1]Chemnitz University of Technology, Germany
[2]InfAI, Leipzig, Germany

**Abstract**

Large Language Models (LLMs) are increasingly used for tasks involving Knowledge Graphs (KGs), whose evaluation typically focuses on accuracy and output correctness. We propose a complementary task characterization approach using three complexity frameworks from cognitive psychology. Applying this to the LLM-KG-Bench framework, we highlight value distributions, identify underrepresented demands and motivate richer interpretation and diversity for benchmark evaluation tasks.

**Keywords**

Task characterization, Benchmark evaluation, LLM, Knowledge Graph, RDF, SPARQL

## 1. Introduction

Large Language Models (LLMs) are increasingly applied to structured knowledge tasks involving query generation, data interpretation and interaction with Knowledge Graphs (KGs) [1, 2]. As a result, the LLM-KG-Bench has been introduced [3, 4] that aim to assess model performance in KG-related context in an automated way. Evaluation for such benchmark tasks typically focuses on correctness and surface-level output features, but provide limited insight into the deeper nature of the tasks themselves, specifically, what kinds of knowledge and operations they demand. In this paper, we propose a task characterization framework for evaluation benchmarks using cognitive complexity frameworks. These allow us to describe each task in terms of the minimal operational and structural requirements expected for successful task completion. Our aim is to support deeper understanding of task diversity and complexity and complement performance-oriented evaluation with structure insight, extending our previous work [5].

## 2. Background and Related Work

Understanding the difficulty and structure of tasks often requires going beyond surface-level features. In cognitive science and educational research, several frameworks have been developed to describe the complexity of tasks based on the type of knowledge involved and the mental operations required. One of the most well-known is *Bloom's Taxonomy* [6], which was originally developed for classification of educational goals based on the required cognitive complexity level. The taxonomy is grounded on behavioral observations of learning processes and classifies cognitive processes from simple recall to higher-level reasoning and creative generation. A revision was made in order to better fit modern views of cognitive psychology [7], which also introduced a complementary dimension. The new *Knowledge Dimension* distinguishes between types of knowledge required for completing different tasks. In parallel, *Relational Complexity Theory* [8] originates from developmental and comparative psychology and draws the notion of relational arity from formal systems in logic and computer science, including relational

*SEMANTiCS'25: International Conference on Semantic Systems, September 3–5, 2025, Vienna, Austria*
*Corresponding author.
sara.todorovikj@informatik.tu-chemnitz.de (S. Todorovikj)
0000-0002-2418-1358 (S. Todorovikj); 0000-0001-5260-5181 (L. Meyer); 0000-0003-0762-8688 (M. Martin)

**Table 1**
Overview of cognitive complexity frameworks and possible values.

| Framework | Values |
|---|---|
| Bloom's Taxonomy - Cognitive Processes [6] | Remember, Understand, Apply, Analyze, Evaluate, Create |
| Knowledge Dimensions [7] | Factual, Conceptual, Procedural, Metacognitive |
| Relational Complexity [8] | Low, Medium, High |

database theory. It formalizes task difficulty in terms of the number of entity and relations that must be simultaneously processed.

These frameworks form the basis for our task characterization approach, which we apply to LLM-KG-Bench as an illustration. The LLM-KG-Bench framework was developed to address the lack of scalable evaluation tools for LLMs targeting KG tasks such as RDF serialization, SPARQL query generation and structured extraction. The framework supports a wide range of tasks with built-in correction cycles and output validation, emphasizing automated, reproducible evaluation across a broad selection of models. Here, we provide a short overview and description of the task groups used in LLM-KG-Bench, for more details, see [3, 9, 10, 11, 4, 5].

#### RDF-related Tasks

*FactExtractStatic* Extract facts from a textual fact sheet and create a KG [3, 9].

*RdfConnectionExplainStatic* Find the shortest connection between two nodes in an RDF graph [9, 4].

*RdfFriendCount* Identify the node with the most incoming edges [9, 4].

*RdfSyntaxFixList* Correct a syntactically invalid RDF graph [4].

*TurtleSampleGeneration* Generate small Turtle KGs satisfying given requirements [3, 9].

#### SPARQL-related Tasks

*Sparql2AnswerList* Given a small KG and a *SPARQL SELECT* query, return the respective result set for the query [11].

*Text2AnswerList* Return the result set answering a given textual question on a given KG (withot a *SPARQL SELECT* query) [11].

*Text2SparqlList* Given a KG and its description, construct a *SPARQL SELECT* query corresponding to a given natural language query [11].

*SparqlSyntaxFixingList* Given a *SPARQL SELECT* query with syntax errors, return a corrected query [11].

## 3. Task Characterization

To understand what kinds of abilities and operations are required by benchmarking tasks, we apply structured characterization criteria drawn from the three established frameworks, as introduced above. While we adopt terminology from cognitive psychology, we do not claim that LLMs engage in these processes in a human sense. Rather, we assess the extent to which their outputs reflect behavior consistent with such operations. Table 1 provides an overview of all possible values across the three frameworks. In the following, we describe the interpretation and assignment criteria for each value.

**Table 2**
Characterization of Benchmark Evaluation Tasks, first submitted at [5]

| Task | Cognitive Process | Knowledge Dimension | Relational Complexity |
|---|---|---|---|
| RDF related: | | | |
| FactExtractStatic | Understand, Create | Conceptual, Procedural | Medium |
| RdfConnectionExplainStatic | Understand, Analyze | Conceptual | Medium |
| RdfFriendCount | Apply | Procedural | Low |
| RdfSyntaxFixList | Understand, Apply | Factual, Procedural | Low |
| TurtleSampleGeneration | Understand, Create | Conceptual, Procedural | Medium |
| SPARQL related: | | | |
| Sparql2AnswerList | Understand, Apply | Conceptual, Procedural | Low |
| Text2AnswerList | Understand, Apply | Conceptual, Procedural | Low |
| Text2SparqlList | Understand, Create | Conceptual, Procedural | Low |
| SparqlSyntaxFixingList | Understand, Apply | Factual, Procedural | Low |

### Bloom's Taxonomy - Cognitive Processes

*Remember* The task depends primarily on "mechanically" recalling facts or definitions without further processing.

*Understand* The task requires interpreting given information, structures or queries without fundamentally transforming or generating new representations.

*Apply* A known procedure or pattern must be correctly executed, such as retrieval or following syntactic rules.

*Analyze* The task demands recognizing or decomposing relationships between data, especially when multiple elements or steps must be coordinated.

*Evaluate* A task involves judging the correctness, relevance or quality of a result.

*Create* The task involves generating new content, such as generating queries or data structures.

### Knowledge Dimensions

*Factual* Task execution success depends on recalling or recognizing specific terminology, syntax elements or concrete information.

*Conceptual* Structural or relational understanding is necessary, such as schema structure, data models or logical organization.

*Procedural* The task requires a correct application of known methods, routines or transformation steps.

*Metacognitive* Awareness and control over one's strategies and thinking processes, such as selecting appropriate approaches, planning task execution or monitoring correctness, which might be relevant for more complex or interactive settings.

### Relational Complexity

*Low* The task involves interpreting or manipulating individual binary relations or isolated, simple structures with minimal dependencies.

*Medium* Multiple relations and entities must be processed simultaneously, such as coordinating several triples or variables in a query.

*High* The task involves multiple interrelated entities or nested dependencies that must be simultaneously considered, often requiring more abstract or hierarchical reasoning.

### 3.1. Application to Benchmark Tasks

The assigned values for each task are displayed in Table 2. Note that not all values across the frameworks are represented, as the current set of tasks does not span the full theoretical space. The assigned values represent the *minimal* operational and structural requirements. Some variability in the relational complexity dimension is certainly possible given a prompt that requires more complex operations.

We can observe several recurring value combinations. Most tasks fall into a characterization combining *Understand* and *Apply* as cognitive processes with *Conceptual* and *Procedural* knowledge dimensions and a *Low* level of relational complexity. This reflects the prevalence of tasks requiring interpretation and rule application without substantial structural coordination. Tasks involving generation (*FactExtractStatic*, *TurtleSampleGeneration* and *Text2SparqlList*) are naturally the only ones annotated with the *Create* process. Among them, only the RDF-based generation tasks are assigned *Medium* relational complexity, reflecting the need to coordinate multiple entities and their relationships when constructing a graph. In contrast, SPARQL generation tasks tend to result in single triple pattern and are thereby assigned *Low* relational complexity.

A consistent, expected dependency can be observed between some processes and knowledge types. *Factual* and *Conceptual* knowledge always coincide with *Understand*, as interpreting a meaning inherently involves factual or structural knowledge. On the other hand, *Procedural* knowledge always coincides with *Apply*, *Analyze* or *Create*, since carrying out a certain procedure by definition requires knowing the necessary steps. In the current task set, *Factual* and *Conceptual* knowledge do not co-occur, distinguishing between surface-level terminology and deeper structural comprehension. Similarly, *Apply*, *Analyze* and *Create* do not co-occur, as they all describe mutually exclusive operations that either follow a procedure, decompose a structure, or generate new ones.

## 4. Discussion and Outlook

In this paper we proposed a task characterization that provides a complementary perspective on benchmark design and evaluation beyond accuracy metrics, inspired by theories of cognitive complexity. This can guide the creation of more balanced and targeted benchmarks by ensuring diversity across the different dimensions. Moreover, it enables identification of potential blind spots in model behavior for tasks that require similar processing.

We demonstrate how to assign the characterization values on a set of evaluation tasks from the LLM-KG-Bench framework. Several values do not appear in the task set due to current design preferences, but that does not imply that such dimensions are irrelevant or unassignable. In cognitive processes, we note *Remember* which would describe a task that asks for reproduction of terminology or exact syntax, e.g., listing reserved SPARQL keywords from memory, while *Evaluate* would require making judgments between alternative options, e.g., selecting the most efficient query. One knowledge dimension was not assigned, *Metacognitive* knowledge, which might be tackled by tasks that require justification, such as explaining the reasoning behind a generated query. Finally, *High* relational complexity would emerge in tasks requiring coordination of more than two entity roles simultaneously, like multi-dimensional event data or nested dependencies. This suggests a direction for extending task design to capture a broader range of structural demands.

The proposed framework could be applied to other benchmarks in the semantic web and beyond, allowing for cross-benchmark comparisons of task complexity profiles. It could also be integrated into such evaluation pipelines, helping understand the types of processes the models succeed or struggle with. In turn, this could support more systematic error analysis, design, and task selection.

## Acknowledgments

This work was partially supported by grants from the German Federal Ministry of Education and Research (BMBF) to the projects ScaleTrust (16DTM312D) and KupferDigital2 (13XP5230L).

## Declaration on Generative AI

The authors have not employed any Generative AI tools.

## References


[1] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 3580–3599. doi:10.1109/tkde.2024.3352100.

[2] L.-P. Meyer, C. Stadler, J. Frey, N. Radtke, K. Junghanns, R. Meissner, G. Dziwis, K. Bulert, M. Martin, LLM-assisted knowledge graph engineering: Experiments with ChatGPT, in: C. Zinke-Wehlmann, J. Friedrich (Eds.), First Working Conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow (AITomorrow) 2023, Informatik aktuell, 2024, pp. 103–115. doi:10.1007/978-3-658-43705-3_8.

[3] L.-P. Meyer, J. Frey, K. Junghanns, F. Brei, K. Bulert, S. Gründer-Fahrer, M. Martin, Developing a scalable benchmark for assessing large language models in knowledge graph engineering, in: N. Keshan, S. Neumaier, A. L. Gentile, S. Vahdati (Eds.), Proceedings of the Posters and Demo Track of the 19th International Conference on Semantic Systems (SEMANTICS 2023), volume 3526 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3526/paper-04.pdf.

[4] L.-P. Meyer, J. Frey, D. Heim, F. Brei, C. Stadler, K. Junghanns, M. Martin, LLM-KG-Bench 3.0: A compass for semantic technology capabilities in the ocean of LLMs, in: E. Curry, M. Acosta, M. Poveda-Villalón, M. van Erp, A. Ojo, K. Hose, C. Shimizu, P. Lisena (Eds.), The Semantic Web. ESWC 2025. Lecture Notes in Computer Science, volume 15719, Springer Nature Switzerland, 2025, pp. 280–296. doi:10.1007/978-3-031-94578-6_16.

[5] L.-P. Meyer, J. Frey, F. Brei, D. Heim, S. Gründer-Fahrer, S. Todorovikj2, C. S. Stadler, M. Schröder, N. Arndt, M. Martin, Evaluating large language models for RDF knowledge graph related tasks - the LLM-KG-Bench-Framework 3, Semantic Web (2025). doi:10.5281/zenodo.16779481, submitted for review 05/2025.

[6] B. S. Bloom, Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain, New York: Longman, 1956.

[7] L. W. Anderson, D. R. Krathwohl, A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Addison Wesley Longman, Inc., 2001.

[8] G. S. Halford, W. H. Wilson, S. Phillips, Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology, Behavioral and Brain Sciences 21 (1998) 803–831.

[9] J. Frey, L.-P. Meyer, N. Arndt, F. Brei, K. Bulert, Benchmarking the abilities of large language models for RDF knowledge graph creation and comprehension: How well do LLMs speak turtle?, in: M. Alam, M. Cochez (Eds.), Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG 2023) co-located with the 21th International Semantic Web Conference (ISWC 2023), Athens, November 6-10, 2023, volume 3559 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3559/paper-3.pdf.

[10] J. Frey, L.-P. Meyer, F. Brei, S. Gruender, M. Martin, Assessing the evolution of LLM capabilities for knowledge graph engineering in 2023, in: The Semantic Web: ESWC 2024 Satellite Events, Springer Nature Switzerland, 2025, pp. 51–60. doi:10.1007/978-3-031-78952-6_5.

[11] L.-P. Meyer, J. Frey, F. Brei, N. Arndt, Assessing SPARQL capabilities of large language models, in: E. Vakaj, S. Iranmanesh, R. Stamartina, N. Mihindukulasooriya, S. Tiwari, F. Ortiz-Rodríguez, R. Mcgranaghan (Eds.), Proceedings of the 3rd International Workshop on Natural Language Processing for Knowledge Graph Creation co-located with 20th International Conference on Semantic Systems (SEMANTiCS 2024), volume 3874 of *CEUR Workshop Proceedings*, 2024, pp. 35–53. URL: https://ceur-ws.org/Vol-3874/paper3.pdf.