KGELL: Knowledge Graphs in the Era of Large Language Models

Benchmarking Large Language Model Quality, Reliability, and Trustworthiness: A Review and Catalogue

KGELL Deliverable D1.1

Francesco Osborne, francesco.osborne@open.ac.uk, ORCID: 0000-0001-6557-3131
Matilde Pato, matilde.pato@gmail.com, ORCID: 0000-0001-8976-7651
Nina Hosseini Kivanani, nina.hosseinikivanani@uni.lu, ORCID: 0000-0002-0821-9125
Mariella Dimiccoli, maria.dimiccoli@upc.edu, ORCID: 0000-0002-2669-400X
Roman Šenkeřík, senkerik@utb.cz, ORCID: 0000-0002-5839-4263
Marco Antonio Stranisci, marcoantonio.stranisci@unito.it, ORCID: 0000-0001-9337-7250

Abstract

This deliverable presents a structured catalogue of 88 benchmarks and evaluation metrics relevant to the quality of Large Language Models (LLMs). Each benchmark has been verified against its source publication and classified along eight features identified through Working Group 1 discussion: performance/reasoning (51 benchmarks), KG integration (34), hallucination (18), factfulness (14), structured output quality (13), bias (9), robustness (8), and explainability (4). The catalogue covers domains from general reasoning and commonsense evaluation to biomedicine, climate science, finance, and data science.
The deliverable provides a measurement baseline for the augmentation strategies that subsequent WG1 deliverables will develop. It also supports the wider KGELL Action by clarifying which benchmark families are most relevant for assessing factuality, grounding, structured output, bias, robustness, explainability, and the specific contribution of Knowledge Graphs to LLM evaluation.

This deliverable is based upon work from COST Action KGELL, CA24121, "Knowledge Graphs in the Era of Large Language Models", supported by COST (European Cooperation in Science and Technology).

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

1. INTRODUCTION

Large Language Models (LLMs) have become core components in a growing range of knowledge-intensive applications, from question answering [41, 43] and information extraction [5, 54] to knowledge graph construction [57, 78] and scientific discovery [2, 10, 11].

Their ability to generate fluent text, however, often comes at the cost of factual accuracy: LLMs can produce plausible but incorrect statements (hallucinations), reproduce societal biases encoded in their training data, and fail unpredictably on structured or domain-specific tasks [36]. For applications that depend on reliable, verifiable knowledge, these shortcomings pose a serious obstacle.

The KGELL COST Action (CA24121) investigates how Knowledge Graphs (KGs) [60] can augment LLMs to improve their accuracy, transparency, and trustworthiness [57]. Working Group 1 (WG1) focuses on strategies for using KGs to make LLMs more reliable, including fine-tuning with KG-derived data, retrieval-augmented generation from KGs, and post-processing LLM outputs against KG-based ground truth [65, 77]. A prerequisite for all of these strategies is the ability to *measure* LLM quality along well-defined features: without reliable benchmarks and metrics, it is impossible to assess whether a given KG-augmentation approach actually improves model behaviour.

This deliverable presents a structured catalogue of 88 benchmarks and evaluation metrics relevant to LLM quality. The catalogue was assembled by collecting benchmark contributions from WG1 members, verifying each entry against its source publication, and tagging each benchmark with analysed features using a systematic, paper-grounded classification scheme. Each benchmark is classified along eight features identified through WG1 discussion: performance/reasoning (51 benchmarks), KG integration (34), hallucination (18), factfulness (14), structured output quality (13), bias (9), robustness (8), and explainability (4). Benchmarks may be tagged with more than one feature. The catalogue covers domains from general reasoning to biomedical knowledge extraction, climate science, and trustworthiness assessment. All feature assignments are confirmed based on direct reading of the source papers.

It is important to note that this catalogue is the result of an exploratory, community-driven collection rather than a systematic literature review. The benchmarks included are those contributed and identified by WG1 members during the first months of the Action; the catalogue does not claim exhaustive coverage of the LLM evaluation landscape. Observations about feature distributions and coverage patterns should therefore be interpreted in the context of this sample. A more comprehensive and methodologically systematic analysis is planned for the coming months as the Action progresses.

2. ANALYSED FEATURES

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

2.1 The features

Through discussion within Working Group 1, we identified eight features along which each benchmark in the catalogue is analysed. These features, listed in Table 1, reflect the quality concerns that are most relevant to the KGELL Action's mission and to the current LLM evaluation landscape. Each benchmark is tagged with one or more features based on what its source paper explicitly states it measures; the features are therefore non-exclusive.

KG integration (K) captures benchmarks that evaluate how LLMs interact with knowledge graphs: retrieving information from them, reasoning over their structure, constructing new KGs from text, or producing outputs that conform to ontological schemata. This feature is the most distinctive to the KGELL Action and differentiates the present catalogue from general LLM benchmark surveys. **Factfulness (F)** concerns whether a claim is *true with respect to world knowledge*. It asks "Is this statement factually correct?" and is evaluated against external knowledge sources such as knowledge graphs, encyclopaedias, or verified databases. A factfulness failure is a false statement.

Hallucination (H) concerns whether a claim is *grounded in the input or source material provided to the model*. It asks "Is this statement supported by the context the model was given?" A hallucination failure is an unsupported statement. A model that generates a biography containing invented dates of birth is hallucinating (fabricating unsupported content) regardless of whether those dates happen to be correct; conversely, a model that faithfully reproduces a flawed source document is not hallucinating, even though its output may be factually wrong. The distinction between factfulness and hallucination matters because the two failure modes call for different mitigation strategies: knowledge grounding for factfulness, input-faithfulness constraints for hallucination. In practice, several benchmarks measure both features simultaneously (e.g., CRAG, HaluBench, DefAn), since factual inaccuracy and unsupported generation often co-occur.

Bias (B) targets demographic, societal, or stereotypical bias in model outputs, including stereotype evaluation, toxicity measurement, and fairness assessment.

Performance/Reasoning (P) covers general task accuracy, knowledge breadth, and reasoning ability. Because it spans a wide range of tasks, we further organise it into five thematic sub-groups for the exposition in Section 7: general reasoning and knowledge, graph and KG reasoning, data science and statistical reasoning, mathematical reasoning, and domain-specific evaluation.

Robustness (R) evaluates model behaviour under adversarial inputs, out-of-distribution data, or incomplete and noisy knowledge. It is particularly relevant for KG-augmented systems, where the KG used for augmentation may be partial, outdated, or contain errors.

Structured output quality (Q) measures the correctness of LLM-generated structured formats, including JSON, tables, SPARQL queries, RDF triples, and scene graphs. It is a prerequisite for reliable KG construction: an LLM that cannot produce syntactically valid and ontology-conformant triples cannot populate a KG.

Explainability (E) captures benchmarks that evaluate the quality of explanations or justifications produced by the system. It asks "Can the model articulate *why* it produced a given answer,

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Table 1: The eight analysed features used in this catalogue. Features are non-exclusive: each benchmark may be tagged with one or more. A feature is assigned only when the source paper explicitly states that it measures the corresponding aspect.

#	Code	Description
1	K	**KG integration**: retrieval from KGs, graph reasoning, triple extraction, KG construction, ontology conformance
2	F	**Factfulness**: truthfulness, factual accuracy, alignment with verifiable facts
3	H	**Hallucination**: generating content not grounded in source material or evidence
4	B	**Bias**: demographic, societal, or stereotypical bias in model outputs
5	P	**Performance/Reasoning**: general task accuracy, knowledge breadth, reasoning ability
6	R	**Robustness**: behaviour under adversarial inputs, out-of-distribution data, or incomplete knowledge
7	Q	**Structured output quality**: correctness of generated JSON, tables, SPARQL, triples, or other structured formats
8	E	**Explainability**: ability to produce faithful, grounded explanations or justifications, including rationale generation and evidence attribution

grounded in retrievable evidence?" Where Performance/Reasoning measures answer correctness, Explainability concerns the reasoning trace itself: whether rationales are faithful, supporting evidence correctly identified, and explanations complete and coherent.

2.2 Feature assignment methodology

Each benchmark was tagged with features through the following procedure. First, we obtained the source paper (or, for tool-based entries, the official documentation). Second, we identified the evaluation metrics and task descriptions that the paper uses to characterise what the benchmark measures. Third, we assigned features based strictly on what the paper explicitly states. If a paper does not mention a feature, even when one might plausibly infer it, that feature was not assigned. This conservative approach avoids over-interpreting benchmark scope and ensures that the catalogue reflects verified measurement intent rather than assumed coverage.

2.3 Distribution overview

Table 2 shows how the 88 benchmarks distribute across the eight features. The total number of assignments (151) exceeds the number of benchmarks because features are non-exclusive. Performance/Reasoning (51 benchmarks, 58%) and KG integration (34, 39%) are the most populated features. Hallucination, the third largest with 18 benchmarks, reflects the substantial recent growth in hallucination evaluation research.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Table 2: Distribution of 88 benchmarks across the eight analysed features. Counts exceed 88 because assignments are non-exclusive.

Feature	Code	Count
Performance/Reasoning	P	51
KG integration	K	34
Hallucination	H	18
Factfulness	F	14
Structured output quality	Q	13
Bias	B	9
Robustness	R	8
Explainability	E	4
Total assignments		**151**

3. KG INTEGRATION

KG integration is the analysed feature most distinctive to the KGELL Action. It captures benchmarks that evaluate how well LLMs interact with knowledge graphs, whether by retrieving information from them, reasoning over their structure, constructing new KGs from text, or producing structured outputs that conform to ontological schemata. Thirty-four of the 88 benchmarks (39%) are tagged with this feature, making it the second most populated after Performance/Reasoning. We organise these benchmarks into four groups by task type.

KG-augmented retrieval.. Representative recent benchmarks and evaluation studies for KG-informed RAG include CRAG, GraphRAG-Bench, BRINK, mmRAG, RAG vs. GraphRAG, and FinReflectKG-MultiHop. CRAG [89] provides 4,409 factual QA pairs with mock web and KG search APIs, and its mock KGs cover 2.6 million entities. GraphRAG-Bench [84] evaluates GraphRAG on tasks of increasing difficulty, covering fact retrieval, complex reasoning, contextual summarisation, and creative generation. GraphRAG-Bench (Xiao et al.) [85] instead provides 1,018 college-level questions across 16 disciplines and evaluates graph construction, retrieval, answer generation, and rationale quality. BRINK [100] examines KG-RAG under knowledge incompleteness by removing direct evidence and testing reasoning over alternative paths. mmRAG [86] broadens evaluation to modular multi-modal RAG over text, tables, and KGs, integrating queries from six QA datasets. RAG vs. GraphRAG [24] systematically compares standard RAG with four representative categories of GraphRAG systems across question answering and query-based summarisation tasks. FinReflectKG-MultiHop [6] benchmarks multi-hop financial QA over a temporally indexed KG derived from S&P 100 10-K filings from 2022–2024; its curated subset of 555 QA pairs spans intra-document, inter-year, and cross-company reasoning, and the paper reports about 24% higher correctness and about 84.5% lower token usage for KG-guided evidence retrieval than for text-window retrieval. Reported metrics across these works include accuracy, F1, Hits-style metrics, NDCG, ROUGE, and BERTScore.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Graph reasoning and related knowledge benchmarks.. Several benchmarks and probes test how well LLMs handle graph-structured or graph-linked knowledge. MetaQA [97] evaluates multi-hop QA over a movie-domain KG with 1-hop, 2-hop, and 3-hop reasoning. KG-LLM-Bench [42] evaluates LLM reasoning on textualised KGs across five task types, including retrieval, path-based reasoning, local aggregation, multi-hop aggregation, and global analysis. ProGraph [30] measures graph analysis via code generation using six Python graph libraries, with problems scaling to 10^6 nodes. GLBench [32] provides a comprehensive benchmark for GraphLLM methods in supervised and zero-shot settings across real-world graph datasets. GPT4Graph [23] tests both structural and semantic graph understanding. M3STR [95] uses multimodal knowledge graphs based on FB15K-237 with image extensions to evaluate abstractive visual understanding through count, detection, and completion tasks. Related QA and knowledge benchmarks include KILT [62], grounded in a shared Wikipedia snapshot; Mintaka [71], which provides 20,000 English QA pairs annotated with Wikidata entities and translated into eight additional languages; and LAMA [61], which probes factual and commonsense knowledge in pretrained language models using cloze queries derived from T-REx/Wikidata and ConceptNet. In Mintaka, the best overall English baseline reaches 38% Hits@1, while the best KGQA baseline, Rigel, reaches 20% Hits@1.

KG construction and information extraction.. Representative benchmarks, datasets, and evaluation frameworks study how LLMs can build or populate KGs from text. Text2KGBench [46] benchmarks ontology-driven KG generation and defines triple-level hallucination metrics for subject, relation, and object hallucinations alongside ontology conformance. Zero-shot Biomedical Triplet Extraction [20] evaluates zero-shot biomedical triplet extraction as named entity recognition and relation extraction across 61 biomedical corpora covering 151 entity types. LLM-KG-Bench [44] evaluates KG engineering tasks including syntax correction, fact extraction from plaintext, and synthetic dataset generation. ClimateIE [56] provides climate-specific named entity recognition (12 entity types), relation extraction (9 relation classes), and entity linking to a climate taxonomy, with the goal of supporting structured climate knowledge construction. BiodivNERE [1] provides expert-labelled gold-standard corpora for named entity recognition and relation extraction in the biodiversity domain, making it useful for KG construction from ecological literature.

KG checking, editing, and structured KG output.. Representative benchmarks, datasets, and frameworks address KG checking, knowledge editing, query generation, or structured graph-based output evaluation. BioKGBench [35] benchmarks AI agents on biomedical KG checking, including KGQA over a clinical knowledge graph with roughly 484K entities and scientific claim verification. VLKEB [26] evaluates knowledge editing in vision-language models, measuring reliability, generality, locality, and portability of edited knowledge sourced from multi-modal KGs. SPARQL Select [45] evaluates LLM ability to generate correct SPARQL queries and recover from errors. Chatty-Gen [53] generates dialogue benchmarks from KGs such as DBpedia, YAGO, DBLP, and MAG, and incorporates validation steps to improve factual consistency across intermediate outputs. TSG-Bench [87] evaluates scene graph understanding and generation, treating scene graphs as structured triplet representations. BENCH4T3 [17] provides a framework

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

for creating text-to-triple alignment benchmarks from DBpedia. GraphEval [69] represents LLM outputs as KG triples and checks consistency against source context, identifying hallucinated triples and proposing GraphCorrect for KG-based hallucination correction. XplainLLM [16] provides 24,204 instances for evaluating explanation grounding with ConceptNet and graph neural models, showing that KG-grounded explanations can reduce hallucinations. SPINACH [37] evaluates SPARQL query generation over Wikidata using 320 expert-annotated question–query pairs from the Wikidata "Request a Query" forum, outperforming prior baselines on QALD benchmarks by 10–31% F1. SIB sparql-examples [12] provides over 1,000 natural language questions with SPARQL queries spanning 11 interoperable bioinformatics knowledge graphs, standardised with structured metadata.

Three additional benchmarks use KGs specifically for hallucination evaluation. KGHaluBench [68] uses KG-derived question sets to evaluate both breadth and depth of LLM knowledge while accounting for popularity-related difficulty effects. MultiHal [29] constructs a multilingual, multi-hop hallucination benchmark from Wikidata paths, and shows that KG-RAG improves semantic similarity and hallucination-related evaluation across multiple languages. OKGQA [74] evaluates KG-augmented LLMs on open-ended QA, including an OKGQA-P variant with deliberately perturbed KGs to test robustness under KG contamination.

In our taxonomy, twelve of the 34 benchmarks are co-tagged with Hallucination or Factfulness, and seven with Structured output quality, illustrating the tight coupling between KG tasks and other quality concerns.

4. FACTFULNESS

Factfulness concerns the alignment of model outputs with verifiable external facts. As noted in Section 2, it differs from hallucination in the reference standard used for evaluation: factfulness evaluates whether a claim is *true according to world knowledge* (knowledge graphs, encyclopaedias, verified databases), whereas hallucination assesses whether a claim is *supported by the input context* the model was given. A factfulness failure is a false statement; a hallucination failure is an unsupported one. The two often co-occur, but they are conceptually and practically distinct. Fourteen benchmarks in the catalogue measure factfulness.

TruthfulQA [34] is the most widely cited benchmark in this group. It contains 817 questions across 38 categories designed to elicit "imitative falsehoods," i.e., answers that are false but that a model might produce because similar falsehoods are common in its training data. Evaluation uses both a truthfulness rate and an informativeness score, the latter penalising evasive non-answers. FEVER [76] provides 185,445 claims generated from Wikipedia sentences and classified as Supported, Refuted, or NotEnoughInfo, evaluating both claim-level accuracy and evidence retrieval quality. KILT [62] benchmarks knowledge-intensive tasks (fact checking, QA, entity linking, slot filling) grounded on Wikipedia, with provenance retrieval as a key evaluation component. CRAG [89] evaluates factual QA for RAG systems, with a scoring protocol that penalises hallucinated answers more heavily than missing ones. BioKGBench [35] includes a scientific claim verification task that checks biomedical claims against published papers. VLKEB [26] measures

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

knowledge editing accuracy in vision-language models, evaluating whether factual updates are correctly applied across three metrics: Reliability, Locality, and Generality.

Several additional benchmarks complete this feature. FELM [15] annotates factuality at the text-segment level across diverse domains (world knowledge, science, maths, reasoning), with predefined error types. OKGQA [74] demonstrates the significance of integrating KGs with LLMs to help reduce hallucination, even in circumstances where the KGs are contaminated. FEVEROUS [4] extends FEVER to structured evidence (tables), providing 87,026 verified claims annotated with evidence from Wikipedia sentences and table cells, classified as Supports, Refuted, or Not Enough Info. SimpleQA [82] contains 4,326 short, fact-seeking questions with single indisputable answers, adversarially collected against GPT-4, measuring whether models "know what they know" via correct rate, not-attempted rate, and F-score. HaluBench [67] provides 15,000 context-question-answer triplets across real-world domains (finance, medicine) and focuses solely on intrinsic hallucinations (unfaithful to context), rather than extrinsic hallucinations (contradicting factual reality), since in real-world RAG settings, user-provided documents may contain information that conflicts with external knowledge sources. DefAn [66] evaluates 75,000 prompts across eight domains, measuring three hallucination types: fact-contradicting hallucination, prompt misalignment, and response inconsistency, with factuality defined as "the degree of accuracy and truthfulness of the generated text."

Two further papers represent additional perspectives on LLM benchmarking. LAMA [61] is an early-stage LLM development probing framework for factual and commonsense knowledge in pre-trained language models using queries derived from Google-RE, T-REx, ConceptNet, and SQuAD. FActScore [47] proposes an evaluation scheme that decomposes LLM outputs into atomic facts and verifies each against a knowledge source, reporting the proportion of supported claims.

Key metrics for this feature include truthfulness rate, informativeness, retrieval accuracy, claim verification accuracy, knowledge editing accuracy, FActScore, correct-given-attempted rate, and mean precision at k (P@k, this value is 1 if the object is ranked among the top k results, and 0 otherwise). KGs naturally serve as ground truth for factual verification: CRAG, VLKEB, OKGQA, BioKGBench, and LAMA use KG-structured knowledge as their reference source.

5. HALLUCINATION

Hallucination, as used in this catalogue, refers to the generation of content that is not grounded in the provided source material or available evidence.

Eighteen benchmarks measure hallucination, making it the third most populated feature in the catalogue. Several benchmarks offer general frameworks for detecting hallucinations. Some of these benchmarks break down outputs into verifiable units, which aligns with the factfulness feature. TruthfulQA [34] studies imitative falsehoods across 38 question categories (also tagged F). HaluEval [31] provides 35K hallucinated samples across QA, dialogue, and summarisation. HalluLens [8] introduces a comprehensive hallucination benchmark and a taxonomy distinguish-

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

ing extrinsic hallucination (content inconsistent with training data) from intrinsic hallucination (content inconsistent with source material), with dynamic test-set generation to prevent data leakage. Med-HALT [55] targets the medical domain with reasoning-based and memory-based hallucination tests derived from multinational medical examinations. DecodingTrust [81] includes hallucination as one of eight trustworthiness perspectives (also tagged B and R). FELM [15] provides segment-level factuality annotations across world knowledge, science, maths, and writing (also tagged F). HaluBench [67] provides 15,000 RAG-focused samples across finance, medicine, and other domains (also tagged F). DefAn [66] evaluates 75,000 prompts for fact-contradicting hallucination, prompt misalignment, and response inconsistency across eight domains (also tagged F).

Dedicated RAG-specific hallucination receives dedicated attention. CRAG [89] includes eight categories of questions, including “False Premise” questions that test whether retrieval-augmented systems fabricate answers. RAGTruth [83] contributes nearly 18,000 RAG responses with word-level hallucination annotations, enabling fine-grained detection.

Several benchmarks use KGs or KG-augmented retrieval for hallucination grounding. BioKG-Bench [35] aims to “combat hallucinations” via KG checking. Chatty-Gen [53] is a benchmark generation tool that mitigates hallucination in KG-grounded dialogue. Text2KGBench [46] defines subject, relation, and object hallucination at the triple evaluation level. KGHaluBench [68] uses KG entity popularity to estimate statistical difficulty in evaluating LLM knowledge breadth and depth. MultiHal [29] extends hallucination evaluation to a multilingual setting using Wikidata KG paths, demonstrating KG-RAG improvements across languages. GraphEval [69] represents LLM outputs as KG triples and checks consistency against source context, identifying specific hallucinated triples and offering a GraphCorrect method for KG-based correction (also tagged K and E). XplainLLM [16] demonstrates that KG-grounded explanations using ConceptNet and Graph Attention Networks reduce hallucinations in commonsense reasoning tasks (also tagged K and E).

Key metrics include hallucination rate, subject/relation/object hallucination scores, fact precision, FActScore, and word-level detection accuracy. The use of KGs or KG-augmented retrieval for grounding confirms that KG-based evaluation is particularly well suited to detecting and quantifying hallucination.

6. BIAS

Bias concerns demographic, societal, or stereotypical bias in model outputs. Nine benchmarks in the catalogue address this feature, covering stereotype evaluation, toxicity measurement, and comprehensive trustworthiness assessment.

Among stereotype and social bias benchmarks, BIG-bench [73] includes six bias tasks covering gender, religion, race/ethnicity, and nationality. StereoSet [48] measures stereotypical bias across four domains using 16,995 test instances. CrowS-Pairs [49] provides 1,508 examples covering nine types of social bias. BBQ [58] evaluates bias in QA across nine social dimensions (age,

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

disability, gender identity, nationality, physical appearance, race/ethnicity, religion, socioeconomic status, sexual orientation), using ambiguous vs. disambiguated contexts; models average 3.4 percentage points higher accuracy when the correct answer aligns with a social bias than when it conflicts with one. WinoBias [98] measures gender bias in coreference resolution, finding a 21.1-point F1 gap between pro-stereotypical and anti-stereotypical entity linking.

BOLD [19] evaluates bias in open-ended language generation using 23,679 prompts across five domains (profession, gender, race, religion, political ideology) and proposes automated metrics for toxicity, psycholinguistic norms, and text gender polarity. RealToxicityPrompts [21] provides 100K sentence-level prompts for evaluating toxic degeneration, finding that pretrained LMs produce toxic text even from innocuous prompts. DecodingTrust [81] assesses trustworthiness across eight perspectives, including stereotype bias, toxicity, adversarial robustness, and fairness (also tagged H and R). HELM [33] evaluates bias, fairness, and toxicity alongside accuracy, calibration, robustness, and efficiency across 42 scenarios and 30 LLMs (also tagged P and R).

Although bias is now substantially better represented than in the initial catalogue, none of the nine benchmarks was designed for KG-augmented systems. Whether KG augmentation introduces, amplifies, or mitigates bias remains an open empirical question.

7. PERFORMANCE AND REASONING

Performance/Reasoning is the most populated feature in the catalogue, spanning 51 of 88 benchmarks (58%). This reflects the prevalence of task-based evaluation in current LLM research: most benchmarks measure how well a model performs on a defined task, even when other features (KG integration, structured output quality) are also assessed. To manage this breadth, we organise these benchmarks into five thematic sub-groups. These sub-labels are used for exposition only and do not appear as formal feature codes in the catalogue metadata.

7.1 General reasoning and knowledge

Nine benchmarks evaluate broad reasoning and knowledge capabilities. MMLU [25] tests multitask accuracy across 57 subjects with 15,908 multiple-choice questions. BIG-bench [73] provides 204 diverse tasks spanning linguistics, common-sense reasoning, biology, physics, and software development. HellaSwag [91] uses adversarially filtered sentence completions to measure commonsense NLI, with a notable gap between human performance ($>$95%) and early model performance ($<$48%). ARC [18] contains 7,787 grade-school science questions, with a Challenge Set designed to defeat retrieval and co-occurrence baselines. ANLI [51] collects NLI examples through iterative adversarial human-and-model-in-the-loop annotation (HAMLET), with 169,265 examples across three rounds, comprising 162,865 training examples, 3,200 development examples, and 3,200 test examples. SuperGLUE [80] provides eight challenging NLU tasks (boolean QA, causal reasoning, reading comprehension, coreference resolution, word sense disambiguation) as a harder successor to GLUE; the human baseline is approximately 89.8. HELM [33] evaluates 30 LLMs across 42 scenarios on seven metrics, including accuracy, cali-

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

bration, robustness, fairness, bias, toxicity, and efficiency (also tagged B and R). MT-Bench [99] evaluates multi-turn instruction-following using LLM-as-judge methodology, achieving over 80% agreement with human preferences. These benchmarks serve as established baselines against which newer, more targeted benchmarks are often compared. HotpotQA [90] provides 113,000 Wikipedia-based QA pairs requiring multi-hop reasoning, with sentence-level supporting fact annotations; it evaluates both answer accuracy (EM, F1) and explanation quality (Joint F1), covering comparison and bridge-entity reasoning types (also tagged E).

7.2 Graph and KG reasoning

Sixteen benchmarks evaluate reasoning over graph-structured data. This sub-group overlaps substantially with the KG integration feature (Section 3): all 16 benchmarks are also tagged K. The benchmarks here include KILT, MetaQA, GraphRAG-Bench, KG-LLM-Bench, BRINK, ProGraph, GLBench, GPT4Graph, mmRAG, GraphRAG-Bench (Xiao et al.), RAG vs. GraphRAG, Zero-shot Biomedical Triplet Extraction, M3STR, Mintaka, SPINACH, and FinReflectKG-MultiHop. Rather than repeat their descriptions, we refer the reader to Section 3, where these are discussed in detail. From a performance perspective, the key observation is that LLM accuracy on graph reasoning tasks remains considerably lower than on standard text-based benchmarks: ProGraph reports the best accuracy of approximately 36%, and BRINK shows that performance degrades substantially when KG triples are removed.

7.3 Data science and statistical reasoning

Eleven benchmarks evaluate LLMs as data science assistants. StatEval [39] is the first comprehensive statistics benchmark, with 13,817 foundational tasks and 2,374 research-level proof tasks (GPT-4 achieves $<$57% on the latter). StatQA [101] tests statistical task formulation and hypothesis testing applicability across 11,623 examples (best accuracy given by GPT-4 approximately 65%). QRData [38] evaluates data-based statistical and causal reasoning, requiring models to reason over real datasets (GPT-4 achieves approximately 58%). DSCodeBench [72] contains 1,000 data science code generation problems from GitHub (best pass@1 approximately 0.20 given by GPT-4o). DA-Code [27] provides 500 agent-based data science tasks requiring multi-step reasoning (best accuracy approximately 31%). LLM4DS [50] evaluates four LLM-based assistants on 100 StrataScratch problems. DSEval [96] introduces a multi-facet evaluation paradigm for data science agent workflows. BLADE [22] tests agents on open-ended scientific data analysis using 12 expert-annotated datasets. AIRepr [92] measures reproducibility of LLM-generated data science workflows and finds that reproducible solutions are significantly more accurate (64.4% vs. 53.5%). AssistedDS [40] evaluates how external domain knowledge affects LLM performance on tabular prediction tasks. HardML [64] provides 100 challenging ML knowledge questions (SOTA models achieve approximately 30% error rate).

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

7.4 Mathematical reasoning

Four benchmarks focus on mathematical problem-solving. “What Are the Odds?” [59] proposes three benchmark tasks to evaluate probabilistic reasoning of LLMs: estimating percentiles, drawing samples, and calculating probabilities. MathArena [7] is a platform for evaluation of LLMs on the latest math competitions and olympiads (AIME, HMMT, USAMO, IMO) with 162 problems; top models score below 40% on proof-based IMO problems. U-Math [3] is a university-level benchmark designed to evaluate LLMs on complex mathematical reasoning. It covers 1,100 university-level problems across six mathematics subject areas, with 20% requiring visual reasoning; it also introduces μ-MATH, a meta-evaluation suite for assessing LLM judges. RealMath [93] assesses LLMs’ abilities on authentic mathematical tasks extracted from research-level mathematics from arXiv papers (633 QA pairs from approximately 9,000 papers).

7.5 Domain-specific evaluation

Eleven benchmarks evaluate LLMs on domain-specific tasks. In the climate domain, ClimaBench [63] evaluates domain-specific BERT models on seven climate NLP text classification datasets. ClimateEval [28] extends this with 25 tasks across 13 datasets, covering stance detection, claim verification, and misinformation detection. ClimateIE [56] evaluates climate-specific NER and relation extraction (also tagged K). Climate Change NER [9] is a manually curated English-language dataset containing 534 abstracts of climate-related papers with 13 climate entity types as part of the Effective and Efficient Language Models for Scientific Applications (INDUS) suite. In biomedical and clinical NLP, CaseReportBench [94] benchmarks dense information extraction from 138 expert-annotated clinical case reports covering inborn errors of metabolism. ClinBench [79] evaluates 11 LLMs on clinical information extraction across three datasets (TCGA lung cancer staging, MIMIC-IV-ECG atrial fibrillation detection, MIMIC SDOH extraction). Med-HALT [55] provides reasoning-based and memory-based hallucination tests derived from multinational medical examinations. TSG-Bench [87] evaluates scene graph understanding and generation. In biodiversity, BiodivNERE [1] provides gold-standard corpora for NER and relation extraction from biodiversity literature, manually labelled by domain experts. pdfQA [70] benchmarks document-level QA over PDFs with 4,000 QA pairs differentiated across 10 complexity dimensions. In the financial domain, FinTextQA [14] provides 1,262 source-attributed QA pairs extracted from finance textbooks and government agency websites, benchmarking long-form financial question answering with ROUGE, BLEU, GPT-4 scoring, and human evaluation.

8. ROBUSTNESS

Robustness captures behaviour under adversarial inputs, out-of-distribution data, or incomplete knowledge. Eight benchmarks evaluate this feature. BIG-bench [73] studies model brittleness under task reformulation across its 204-task suite. HellaSwag [91] uses adversarial filtering to create examples that systematically expose model weaknesses in commonsense reasoning.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

ANLI [51] iteratively collects human-generated adversarial NLI examples that defeat current SOTA models. BRINK [100] removes direct triples from a KG and measures whether KG-RAG methods can still reason over alternative paths, finding substantial performance degradation. AssistedDS [40] assesses whether LLMs can resist adversarial domain knowledge hints in tabular data to evaluate their ability to critically identify harmful domain knowledge, finding that models uncritically adopt adversarial hints and suffer severe performance degradation.

Three more benchmarks contribute to this feature. OKGQA [74] benchmarks LLMs in open-ended question answering and includes the OKGQA-P variant that deliberately perturbs KG semantics and structure to test model robustness under KG contamination. DecodingTrust [81] assesses model trustworthiness through adversarial robustness, out-of-distribution robustness, and robustness on adversarial demonstrations, among its eight perspectives. HELM [33] includes robustness as one of seven evaluation metrics across 42 scenarios (also tagged P and B).

Robustness coverage has improved but remains limited for KG-specific settings. BRINK and OKGQA-P are the only benchmarks that test KG-augmented systems under knowledge incompleteness or contamination.

9. STRUCTURED OUTPUT QUALITY

Structured output quality measures the correctness of LLM-generated structured formats, including JSON, tables, SPARQL queries, RDF triples, and scene graphs. Thirteen benchmarks address this feature. The group divides naturally into two clusters: general-purpose structured output benchmarks and KG-specific structured output benchmarks.

Among general-purpose benchmarks, StructEval [88] evaluates generation across 18 formats (JSON, YAML, CSV, XML, HTML, React, SVG, LaTeX, TikZ, and others) and 44 task types, measuring syntax validity, keyword matching, and visual quality. LLMStructBench [75] focuses on JSON extraction from natural language, with 995 examples and metrics for token-level F1 and document-level correctness. ScrapeGraphAI-100k [13] provides 93,695 real-world web extraction examples, evaluating JSON schema conformance and value accuracy. Table Extraction [52] compares multimodal LLMs with traditional deep learning for table extraction from images, measuring structural layout (GriTS F1) and text content accuracy. CaseReportBench [94] and ClinBench [79] evaluate structured extraction in clinical settings.

Among KG-specific benchmarks, SPARQL Select [45] measures correctness and error recovery on SELECT queries. Text2KGBench [46] evaluates the ability of language models to generate KGs from natural language text guided by an ontology. LLM-KG-Bench [44] tests RDF/Turtle syntax correctness and fact extraction quality. TSG-Bench [87] evaluates LLMs' ability to understand scene graphs and generate them from textual narratives. BENCH4T3 [17] is a framework for generating text-to-triple alignment benchmarks for any language using KGs as input. SPINACH [37] is a knowledge-base question answering benchmark that evaluates LLM-based SPARQL query generation over Wikidata with 320 expert-annotated question-query pairs, measuring both exact match and F1 on generated queries (also tagged K and P). SIB sparql-examples [12] provides

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

over 1,000 natural language questions with SPARQL queries across 11 federated bioinformatics KGs, standardised via SHACL metadata and used as a benchmark resource by the ESWC TEXT2SPARQL challenge (also tagged K).

10. CROSS-CUTTING ANALYSIS

The preceding sections presented benchmarks grouped by analysed feature. Here we identify patterns that cut across features.

Feature overlap.. The 88 benchmarks receive 151 feature assignments in total, an average of 1.72 features per benchmark. Many benchmarks serve multiple evaluation purposes simultaneously: CRAG, for instance, is tagged K, F, and H because it evaluates KG-augmented retrieval, factual accuracy, and hallucination within a single protocol. The highest overlap occurs between KG integration and Performance/Reasoning (19 benchmarks share both tags), reflecting the fact that graph reasoning is both a performance task and a KG-specific capability.

Metric families across features.. The same metric family can serve different features depending on context. F1, the most widely used metric in the catalogue, measures factual claim coverage (Factfulness), triple extraction accuracy (KG integration), entity recognition performance (Performance/Reasoning), and structural correctness (Structured output quality). Accuracy similarly spans from knowledge QA (Factfulness) to code execution correctness (Performance/Reasoning).

Intrinsic vs. extrinsic metrics.. Benchmarks in the catalogue use both intrinsic metrics, which evaluate output quality directly (e.g., hallucination rate, ontology conformance, Stereotype Score), and extrinsic metrics, which measure downstream task performance (e.g., QA accuracy, pass@k). Intrinsic metrics are concentrated in the Hallucination, Bias, and Structured output quality features, while Performance/Reasoning relies almost entirely on extrinsic metrics. For KG integration, both types appear: retrieval accuracy and NDCG@K are extrinsic, while triple-level hallucination and ontology conformance are intrinsic.

Temporal and thematic trends.. The catalogue spans benchmarks from 2018 (FEVER, MetaQA, ARC) to 2026 (GraphRAG-Bench, LLMStructBench, RealMath, BENCH4T3). Earlier benchmarks tend to evaluate general capabilities (truthfulness, commonsense reasoning, bias), while more recent ones target specialised tasks: KG-augmented retrieval, structured output generation, data science reasoning, and domain-specific NLP tasks. This shift reflects the maturation of LLM evaluation from broad capability testing towards fine-grained, application-oriented measurement.

Domain coverage.. The majority of benchmarks target general-domain tasks. Domain-specific coverage clusters around five areas: biomedical and clinical NLP (BioKGBench, CaseReportBench, ClinBench, Zero-shot Biomedical Triplet Extraction, Med-HALT), climate and earth science (ClimaBench, ClimateEval, ClimateIE, Climate Change NER), biodiversity (BiodivNERE), finance (FinReflectKG-MultiHop, FinTextQA), and data science (11 benchmarks in the P.c sub-group).

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

11. DOMAIN-SPECIFIC PERSPECTIVES (WORKING GROUP 2)

Working Group 2 approaches the benchmarks and metrics catalogued above from the standpoint of concrete application domains. Where WG1 targets general-purpose measurement, WG2's mandate is to evaluate LLM performance in specific downstream tasks; the eight analysed features of Section 2—and in particular Factfulness, Hallucination, and Bias—therefore act as a specialisation layer that must be instantiated differently for each domain's evidence structures, gold standards, and fairness concerns. The following perspectives, contributed by the WG2 Healthcare and Scientific Research subgroups, illustrate how the general evaluation constructs of this deliverable translate into domain-specific measurement protocols and candidate benchmarks.

11.1 Healthcare

A key challenge that cuts across all identified healthcare tasks is ensuring that LLM outputs are both factually reliable and free from systematic biases, particularly when models are used to support clinical decision-making or biomedical research. Knowledge graphs (KGs) provide a principled framework for addressing these issues by serving as structured, curated sources of biomedical knowledge against which LLM-generated content can be grounded and verified. They can be used both during inference (e.g., retrieval-augmented generation, graph-guided prompting, or constrained reasoning) and during evaluation by verifying whether generated entities, relations, explanations, or recommendations are supported by the KG. Factuality can be quantified using metrics such as entity precision, recall and F1, relation accuracy, triple-level precision/recall, graph consistency, hallucination rate (percentage of unsupported claims), claim verification accuracy, citation accuracy, and semantic entailment scores computed with Natural Language Inference (NLI) models. For tasks involving generated explanations, metrics such as faithfulness, answer groundedness, and attribution precision can assess whether each statement is supported by retrieved evidence from the KG or associated literature. Benchmark datasets with expert-annotated answers or clinical guidelines provide gold standards against which exact-match, ROUGE, BLEU, BERTScore, or embedding-based similarity metrics can be complemented by KG-based verification. Bias should be evaluated across demographic (age, sex, ethnicity), clinical (disease stage, comorbidities), and institutional (hospital or cohort) dimensions using fairness metrics such as demographic parity difference, equal opportunity difference, equalized odds, predictive parity, calibration within groups, and subgroup-specific performance measures (e.g., AUROC, F1, sensitivity, specificity). Counterfactual fairness can be assessed by modifying protected attributes while keeping the remaining clinical context fixed and measuring the stability of model predictions. Finally, robustness can be evaluated through adversarial prompting, perturbation analysis, uncertainty estimation, and consistency across repeated generations. Together, these approaches provide a comprehensive evaluation framework in which biomedical KGs not only enhance LLM reasoning but also enable systematic measurement of factual correctness, explainability, robustness, and fairness across a wide range of healthcare applications.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

11.2 Scientific Research

In scientific research, the central question is not only whether an LLM gives a factual answer, but whether access to a knowledge graph makes that answer more grounded, traceable and reliable than a strong text-only system. The main failures are familiar: fabricated references and DOIs, real citations attached to the wrong claim, methods or results attributed to the wrong paper, reversed findings, and syntheses that sound plausible but are not supported by the cited evidence. KGs are relevant here because science already has graph-like structure: publications, authors, citation links, entities, methods, datasets, ontologies and, in some settings, structured claims. A KG can help by grounding entities into controlled vocabulary, connecting claims to citation and evidence trails, supporting multi-hop retrieval, and making the relations between claims visible. Graph value depends on coverage, entity resolution, ontology quality, provenance and taxonomy. In addition, knowledge graphs may support scientific reproducibility and transparency by exposing provenance chains, preserving links between claims, methods, datasets and publications, and enabling explicit representation of uncertainty, conflicting findings and disciplinary context. For this domain, metrics should therefore compare LLM-only, text-RAG and graph-supported settings under matched conditions, measuring citation resolution, evidence-span support, entity grounding, result/verdict correctness, source and language coverage, graph-path validity, whether the system recognises insufficient evidence, and how much the graph improves on a strong text-only baseline. Candidate benchmarks include SciFact or HealthVer for claim verification (the most direct factfulness test), QASPER or SciQA for scholarly question answering, and SciCite for citation-intent classification and misattribution.

12. CONCLUSION

This deliverable has presented a structured catalogue of 88 benchmarks relevant to LLM quality evaluation, all with confirmed feature assignments based on direct reading of their source publications. The benchmarks are analysed along eight features identified through Working Group 1 discussion: performance/reasoning (51 benchmarks), KG integration (34), hallucination (18), factfulness (14), structured output quality (13), bias (9), robustness (8), and explainability (4).

The catalogue's most distinctive contribution is the KG integration feature, which covers 34 benchmarks spanning KG-augmented retrieval, graph reasoning, KG construction, and KG checking. This feature reflects the KGELL Action's core mission and differentiates the present catalogue from general LLM benchmark surveys. Cross-cutting analysis reveals substantial overlap between features (151 assignments across 88 benchmarks, an average of 1.72 per benchmark), with KG integration and Performance/Reasoning sharing the largest intersection (19 benchmarks).

As noted in Section 1, this catalogue represents an initial, community-driven collection assembled from contributions by WG1 members during the first months of the Action. It does not claim exhaustive coverage of the LLM evaluation landscape. In the coming months, we plan to extend

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

this work into a comprehensive, methodologically systematic survey following established review protocols, with the aim of producing a more complete picture of the evaluation resources available for KG-augmented LLM systems. We encourage continued community contributions to support this effort.

This deliverable provides a measurement baseline for the KGELL Action. The augmentation strategies developed in subsequent WG1 deliverables (D1.2, D1.3, D1.4) can be assessed against the benchmarks and metrics catalogued here. The evaluation frameworks that WG6 will develop (D6.1–D6.3) can build on the feature definitions and analysis presented in this deliverable.

A. BENCHMARK CATALOGUE

Table 3 lists all 88 benchmarks in the catalogue, sorted by catalogue number. Feature codes follow the definitions in Section 2: K = KG integration, F = Factfulness, H = Hallucination, B = Bias, P = Performance/Reasoning, R = Robustness, Q = Structured output quality, E = Explainability.

Table 3: Complete benchmark catalogue (88 entries).

#	Benchmark	Year	Features	Primary Metrics
1	TruthfulQA [34]	2022	F, H	Truthfulness rate, informativeness
2	MMLU [25]	2021	P	Accuracy / Exact Match
3	BIG-bench / BBH [73]	2023	B, P, R	Task-specific accuracy; calibration
4	HellaSwag [91]	2019	P, R	Accuracy
5	ARC [18]	2018	P	Accuracy
6	FEVER [76]	2018	F	Accuracy, F1, Precision, Recall
7	ANLI [51]	2020	P, R	Accuracy
8	StereoSet [48]	2021	B	SS, LMS, ICAT
9	CrowS-Pairs [49]	2020	B	Bias Score
10	KILT [62]	2021	K, F, P	F1, Accuracy, Retrieval Accuracy
11	MetaQA [97]	2018	K, P	Accuracy (by hop count)
12	CRAG [89]	2024	K, F, H	Accuracy, hallucination rate
13	GraphRAG-Bench [84]	2026	K, P	Accuracy (retrieval, reasoning, summarisation)
14	KG-LLM-Bench [42]	2025	K, P	Accuracy (retrieval, reasoning, aggregation)
15	SPARQL Select [45]	2024	K, Q	Query correctness
16	BioKGBench [35]	2024	K, F, H	KGQA accuracy, claim verification F1
17	BRINK [100]	2026	K, P, R	Hits@Any, Precision, Recall, F1
18	ProGraph [30]	2024	K, P	Pass rate, Accuracy
19	GLBench [32]	2024	K, P	Accuracy, Macro-F1
20	GPT4Graph [23]	2023	K, P	Accuracy (node classif., graph classif., KGQA)

Continued on next page

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Table 3 continued

#	Benchmark	Year	Features	Primary Metrics
21	mmRAG [86]	2025	K, P	NDCG@K, MAP@K, Hits@K, EM, F1
22	GraphRAG-Bench (Xiao) [85]	2025	K, P, E	Answer accuracy, rationale quality
23	RAG vs. GraphRAG [24]	2025	K, P	F1, Accuracy, ROUGE-2, BERTScore
24	Chatty-Gen [53]	2025	K, H	Success rate, error counts
25	TSG-Bench [87]	2025	K, P, Q	Accuracy, Precision, Recall, F1
26	CaseReportBench [94]	2025	P, Q	TSR%, BLEU, ROUGE-L
27	StructEval [88]	2025	Q	Syntax, Keyword, VQA scores
28	LLMStructBench [75]	2026	Q	Token-level F1, DOC
29	ClinBench [79]	2025	P, Q	F1, Precision, Recall
30	ScrapeGraphAI-100k [13]	2026	Q	Precision, Recall, F1, JSON Validity
31	Table Extraction [52]	2025	Q	GriTS F1, Structural Layout F1
32	BENCH4T3 [17]	2026	K, Q	ROUGE, BLEU, cosine similarity, NDCG
33	Zero-shot Bio. Triplets [20]	2025	K, P	NER F1, RE F1
34	LLM-KG-Bench [44]	2023	K, Q	Accuracy, error rates
35	Text2KGBench [46]	2023	K, H, Q	Precision, Recall, F1
36	StatEval [39]	2025	P	Computational correctness, proof accuracy
37	StatQA [101]	2024	P	Accuracy
38	QRData [38]	2024	P	Accuracy
39	DSCodeBench [72]	2026	P	pass@k
40	DA-Code [27]	2024	P	Accuracy
41	LLM4DS [50]	2024	P	Execution correctness, similarity score
42	DSEval [96]	2024	P	Pass rate, error propagation rate
43	BLADE [22]	2024	P	Precision, Coverage@k, F1
44	AIRepr [92]	2025	P	Reproducibility, accuracy
45	AssistedDS [40]	2025	P, R	Submission validity, hint recall, predictive accuracy
46	HardML [64]	2024	P	Accuracy
47	“What Are the Odds?” [59]	2024	P	Estimation/calculation accuracy
48	MathArena [7]	2025	P	Competition-level accuracy
49	U-Math [3]	2024	P	Accuracy
50	RealMath [93]	2025	P	Accuracy
51	ClimaBench [63]	2025	P	Classification accuracy/F1
52	ClimateEval [28]	2025	P	Accuracy, F1
53	BiodivNERE [1]	2022	K, P	F1, Precision, Recall
54	ClimateIE [56]	2025	K, P	F1, Precision, Recall
55	Climate Change NER [9]	2024	P	NER F1
56	pdfQA [70]	2026	P	G-Eval correctness score
57	M3STR [95]	2025	K, P	Visual reasoning accuracy

Continued on next page

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Table 3 continued

#	Benchmark	Year	Features	Primary Metrics
58	VLKEB [26]	2024	K, F	Knowledge editing accuracy
59	FActScore [47]	2023	F, H	FActScore
60	HaluEval [31]	2023	H	Hallucination detection accuracy
61	Med-HALT [55]	2023	H, P	Hallucination rate, reasoning accuracy
62	FELM [15]	2023	F, H	Factuality detection accuracy
63	RAGTruth [83]	2024	H	Hallucination detection accuracy
64	HalluLens [8]	2025	H	Hallucination detection accuracy
65	KGHaluBench [68]	2026	K, H	Breadth accuracy, depth accuracy
66	MultiHal [29]	2025	K, H	Semantic similarity, NLI
67	OKGQA [74]	2025	K, F, R	QA accuracy, hallucination rate
68	BBQ [58]	2022	B	Bias Score
69	WinoBias [98]	2018	B	F1 difference (pro- vs. anti-stereo.)
70	BOLD [19]	2021	B	Toxicity, sentiment scores
71	DecodingTrust [81]	2023	H, B, R	Multi-perspective trust scores
72	RealToxicityPrompts [21]	2020	B	Toxicity score (Perspective API)
73	HELM [33]	2023	B, P, R	Accuracy, calibration, robustness, fairness, bias, toxicity, efficiency
74	MT-Bench [99]	2023	P	LLM judge agreement
75	SuperGLUE [80]	2019	P	Accuracy/F1, SuperGLUE score
76	HaluBench [67]	2024	F, H	Hallucination detection accuracy
77	DefAn [66]	2024	F, H	Factual hallucination rate
78	FEVEROUS [4]	2021	F	Evidence + verdict accuracy
79	SimpleQA [82]	2024	F	Correct rate, F-score
80	Mintaka [71]	2022	K, P	Hits@1
81	HotpotQA [90]	2018	P, E	EM, F1, Joint F1
82	GraphEval [69]	2024	K, H, E	Balanced accuracy, ROUGE
83	LAMA [61]	2019	K, F	P@1
84	XplainLLM [16]	2024	K, H, E	Debugger-score
85	SPINACH [37]	2024	K, P, Q	EM, F1
86	SIB sparql-examples [12]	2025	K, Q	NL–SPARQL pair count, federated coverage
87	FinReflectKG-MultiHop [6]	2025	K, P	Correctness, Token utilisation
88	FinTextQA [14]	2024	P	ROUGE, BLEU, GPT-4 score

B. BENCHMARK LINKS

Table 4 provides official links to benchmark resources (code repositories, project pages, or dataset downloads). Where no dedicated project page was found, the link points to the paper on arXiv or the publication venue. A dash indicates that no public link could be identified at the time of writing.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Table 4: Official links for all 88 benchmarks.

#	Benchmark	Official Link
1	TruthfulQA	https://github.com/sylinrl/TruthfulQA
2	MMLU	https://github.com/hendrycks/test
3	BIG-bench / BBH	https://github.com/google/BIG-bench
4	HellaSwag	https://github.com/rowanz/hellaswag
5	ARC	https://github.com/allenai/ai2_arc
6	FEVER	https://github.com/awslabs/fever
7	ANLI	https://github.com/facebookresearch/anli
8	StereoSet	https://github.com/moinnadeem/StereoSet
9	CrowS-Pairs	https://github.com/nyu-mll/crows-pairs
10	KILT	https://github.com/facebookresearch/KILT
11	MetaQA	https://github.com/yuyuz/MetaQA
12	CRAG	https://github.com/facebookresearch/CRAG
13	GraphRAG-Bench	https://github.com/GraphRAG-Bench/GraphRAG-Benchmark
14	KG-LLM-Bench	https://arxiv.org/abs/2504.07087
15	SPARQL Select	https://github.com/AKSW/LLM-Sparql-Results-2024
16	BioKGBench	https://github.com/westlake-autolab/BioKGBench
17	BRINK	https://aclanthology.org/2026.eacl-long.114/
18	ProGraph	https://github.com/BUPT-GAMMA/ProGraph
19	GLBench	https://github.com/NineAbyss/GLBench
20	GPT4Graph	https://arxiv.org/abs/2305.15066
21	mmRAG	https://arxiv.org/abs/2505.11180
22	GraphRAG-Bench (Xiao)	https://github.com/jeremycp3/GraphRAG-Bench
23	RAG vs. GraphRAG	https://arxiv.org/abs/2502.11371
24	Chatty-Gen	https://doi.org/10.1145/3709681
25	TSG-Bench	https://github.com/docworlds/tsg-bench
26	CaseReportBench	https://github.com/cindyzhangxy/CaseReportBench
27	StructEval	https://github.com/TIGER-AI-Lab/StructEval
28	LLMStructBench	https://arxiv.org/abs/2602.14743
29	ClinBench	https://github.com/ismaelvillanuevamiranda/ClinBench
30	ScrapeGraphAI-100k	https://github.com/ScrapeGraphAI/scrapegraph-100k-paper
31	Table Extraction	https://aclanthology.org/2025.xllm-1.2/
32	BENCH4T3	https://github.com/Visot/BENCH4T3
33	Zero-shot Bio. Triplets	https://aclanthology.org/2025.bionlp-1.9/
34	LLM-KG-Bench	https://github.com/AKSW/LLM-KG-Bench
35	Text2KGBench	https://github.com/cenguix/Text2KGBench
36	StatEval	https://stateval.github.io/
37	StatQA	https://github.com/HKUSTDial/StatQA
38	QRData	https://xxxiaol.github.io/QRData/
39	DSCodeBench	https://github.com/ShuyinOuyang/DSCodeBench
40	DA-Code	https://github.com/yiyihum/da-code
41	LLM4DS	https://github.com/DataForScience/LLM4DS
42	DSEval	https://github.com/MetaCopilot/dseval
43	BLADE	https://github.com/behavioral-data/BLADE
44	AIRepr	https://aclanthology.org/2025.findings-emnlp.539/
45	AssistedDS	https://arxiv.org/abs/2506.13992

Continued on next page

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Table 4 continued

#	Benchmark	Official Link
46	HardML	https://arxiv.org/abs/2501.15627
47	“What Are the Odds?”	https://github.com/yahskapar/LLMs-and-Probabilistic-Reasoning
48	MathArena	https://matharena.ai/
49	U-Math	https://toloka.ai/math-benchmark
50	RealMath	https://github.com/ethz-spylab/RealMath
51	ClimaBench	https://doi.org/10.1007/s42452-025-07740-5
52	ClimateEval	https://aclanthology.org/2025.climatenlp-1.13/
53	BiodivNERE	https://zenodo.org/records/6575865
54	ClimateIE	https://aclanthology.org/2025.climatenlp-1.6/
55	Climate Change NER	https://arxiv.org/abs/2405.10725
56	pdfQA	https://github.com/tobischimanski/pdfQA
57	M3STR	https://github.com/zjukg/M3STR
58	VLKEB	https://github.com/VLKEB/VLKEB
59	FActScore	https://github.com/shmsw25/FActScore
60	HaluEval	https://github.com/RUCAIBox/HaluEval
61	Med-HALT	https://github.com/medhalt/medhalt
62	FELM	https://github.com/hkust-nlp/felm
63	RAGTruth	https://github.com/ParticleMedia/RAGTruth
64	HalluLens	https://github.com/facebookresearch/HalluLens
65	KGHaluBench	https://github.com/c0037654Newcastle/KGHaluBench
66	MultiHal	https://arxiv.org/abs/2505.14101
67	OKGQA	https://github.com/Y-Sui/OKGQA
68	BBQ	https://github.com/nyu-mll/BBQ
69	WinoBias	https://github.com/uclanlp/corefBias
70	BOLD	https://github.com/amazon-science/bold
71	DecodingTrust	https://github.com/AI-secure/DecodingTrust
72	RealToxicityPrompts	https://github.com/allenai/real-toxicity-prompts
73	HELM	https://github.com/stanford-crfm/helm
74	MT-Bench	https://github.com/lm-sys/FastChat
75	SuperGLUE	https://super.gluebenchmark.com
76	HaluBench	https://huggingface.co/datasets/PatronusAI/HaluBench
77	DefAn	https://github.com/ashikiut/DefAn
78	FEVEROUS	https://github.com/Raldir/FEVEROUS
79	SimpleQA	https://github.com/openai/simple-evals
80	Mintaka	https://github.com/amazon-science/mintaka
81	HotpotQA	https://hotpotqa.github.io
82	GraphEval	https://arxiv.org/abs/2407.10793
83	LAMA	https://github.com/facebookresearch/LAMA
84	XplainLLM	https://github.com/chen-zichen/XplainLLM_dataset
85	SPINACH	https://github.com/stanford-oval/spinach
86	SIB sparql-examples	https://github.com/sib-swiss/sparql-examples
87	FinReflectKG-MultiHop	https://github.com/finreflectkg/FinReflectKG-MultiHop
88	FinTextQA	https://github.com/AlexJJJChen/FinTextQA

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

REFERENCES

[1] Nora Abdelmageed, Felicitas Löffler, Leila Feddoul, Alsayed Algergawy, Sheeba Samuel, and Jitendra Gaikwad. Biodivnere: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. *Biodiversity Data Journal, 10, e89481*, 2022. doi: 10.3897/bdj.10.e89481.

[2] Tanay Aggarwal, Angelo Salatino, Francesco Osborne, and Enrico Motta. Large language models for scholarly ontology generation: An extensive analysis in the engineering field. *Information Processing & Management*, 63(1):104262, 2026.

[3] Toloka AI and Gradarius. U-math: A university-level benchmark for evaluating mathematical skills in llms. 2024.

[4] Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, et al. Feverous: Fact extraction and verification over unstructured and structured information. In *NeurIPS 2021 (Datasets and Benchmarks Track)*, 2021.

[5] Simone Angioni, Sergio Consoli, Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, and Angelo Salatino. Exploring environmental, social, and governance (esg) discourse in news: An ai-powered investigation through knowledge graph analysis. *IEEE Access*, 12:77269–77283, 2024.

[6] Abhinav Arun, Reetu Raj Harsh, Bhaskarjit Sarmah, and Stefano Pasquali. FinReflectKG – MultiHop: Financial QA benchmark for reasoning with knowledge graph evidence. 2025.

[7] Mislav Balunović, Jasper Dekoninck, Ivo Petrov, and Nikola Jovanović. Matharena: Evaluating llms on uncontaminated math competitions. 2025.

[8] Yejin Bang, Ziwei Ji, Alan Schelten, A. Hartshorn, T. Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. Hallulens: Llm hallucination benchmark. In *ACL 2025*, 2025.

[9] Bishwaranjan Bhattacharjee, Aashka Trivedi, Masayasu Muraoka, et al. Indus: Effective and efficient language models for scientific applications. In *EMNLP 2024 Industry Track*, 2024. doi: 10.18653/v1/2024.emnlp-industry.9.

[10] Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280, 2023.

[11] Francisco Bolanos, Angelo Salatino, Francesco Osborne, and Enrico Motta. Artificial intelligence for literature reviews: opportunities and challenges. *Artificial Intelligence Review*, 57(10):259, 2024.

[12] Jerven Bolleman, Vincent Emonet, Adrian Altenhoff, et al. A large collection of bioinformatics question–query pairs over federated knowledge graphs: methodology and applications. *GigaScience*, 2025. doi: 10.1093/gigascience/giaf045.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

[13] William Brach et al. Scrapegraphai-100k: A large-scale benchmark for web extraction. 2026.

[14] Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. FinTextQA: A dataset for long-form financial question answering. In *ACL 2024*, 2024. doi: 10.18653/v1/2024.acl-long.328.

[15] Shiqi Chen, Yiran Zhao, Jinghan Zhang, Ethan Chern, Siyang Gao, Pengfei Liu, and Junxian He. Felm: Benchmarking factuality evaluation of large language models. In *NeurIPS 2023*, 2023.

[16] Zichen Chen, Jianda Chen, Ambuj Singh, and Misha Sra. Xplainllm: A knowledge-augmented dataset for reliable grounded explanations in llms. In *EMNLP 2024*, 2024.

[17] Víctor Jesús Sotelo Chico, André Gomes Regino, and Júlio Cesar dos Reis. Bench4t3: A framework to create benchmarks for text-to-triples alignment generation. *Journal of the Brazilian Computer Society*, 32(1):85–101, 2026. doi: 10.5753/jbcs.2026.5809.

[18] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. 2018.

[19] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *FAccT 2021*, 2021.

[20] Frederik Gade, Ole Lund, and Marie Lisandra Mendoza. Benchmarking zero-shot biomedical relation triplet extraction across language model architectures. In *BioNLP 2025 (at ACL)*, 2025.

[21] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *EMNLP 2020 Findings*, 2020.

[22] Ken Gu, Ruoxi Shang, Ren Jiang, Keying Kuang, Ren Lin, et al. Blade: Benchmarking language model agents for data-driven science. In *Findings of EMNLP 2024*, 2024. doi: 10.18653/v1/2024.findings-emnlp.815.

[23] Jiayan Guo et al. Gpt4graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. 2023.

[24] Hwan-Soo Han et al. Rag vs. graphrag: A systematic evaluation and key insights. 2025.

[25] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR 2021*, 2021.

[26] Han Huang, Qiang Liu, Tieniu Tan, Liang Wang, et al. Vlkeb: A large vision-language model knowledge editing benchmark. In *NeurIPS 2024*, 2024. doi: 10.52202/079017-0294.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

[27] Yiming Huang, Jianwen Luo, Yan Yu, Yitong Zhang, Fangyu Lei, et al. Da-code: Agent data science code generation benchmark for large language models. In *EMNLP 2024*, 2024. doi: 10.18653/v1/2024.emnlp-main.748.

[28] Murathan Kurfali, Shorouq Zahra, Joakim Nivre, and Gabriele Messori. Climateeval: A comprehensive benchmark for nlp tasks related to climate change. In *ClimateNLP Workshop @ ACL 2025*, 2025. doi: 10.18653/v1/2025.climatenlp-1.13.

[29] Ernests Lavrinovics, Russa Biswas, Katja Hose, and Johannes Bjerva. Multihal: Multilingual dataset for knowledge-graph grounded evaluation of llm hallucinations. 2025.

[30] Li et al. Can large language models analyze graphs like professionals? a benchmark, datasets and models. In *NeurIPS 2024*, 2024.

[31] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *EMNLP 2023*, 2023.

[32] Yuhan Li et al. Glbench: A comprehensive benchmark for graph with large language models. In *NeurIPS 2024*, 2024. doi: 10.52202/079017-1340.

[33] Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*, 1525:140–146, 2023.

[34] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL 2022*, 2022. doi: 10.18653/v1/2022.acl-long.229.

[35] Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z Li, and Kaicheng Yu. Biokgbench: A knowledge graph checking benchmark of ai agent for biomedical science. *arXiv preprint arXiv:2407.00466*, 2024.

[36] Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. Towards trustworthy llms: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57(9):1, 2024.

[37] Shicheng Liu, Sina J. Semnani, Harold Triedman, et al. Spinach: Sparql-based information navigation for challenging real-world questions. In *Findings of EMNLP 2024*, 2024.

[38] Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. In *ACL 2024*, 2024.

[39] Yucheng Lu, Run Yang, Yichen Zhang, Shuguang Yu, et al. Stateval: A comprehensive benchmark for large language models in statistics. 2025.

[40] An Luo, Xun Xian, Du Jin, Fangqiao Tian, Ganghua Wang, et al. Assistedds: Benchmarking how external domain knowledge assists llms in automated data science. In *Findings of EMNLP 2025*, 2025. doi: 10.18653/v1/2025.findings-emnlp.979.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

[41] Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. Unifying large language models and knowledge graphs for question answering: Recent advances and opportunities. In *EDBT*, pages 1174–1177, 2025.

[42] Elan Markowitz, Anil Ramakrishna, and Aram Galstyan. Kg-llm-bench: A scalable benchmark for evaluating llm reasoning on textualized knowledge graphs. 2025.

[43] Antonello Meloni, Diego Reforgiato Recupero, Francesco Osborne, Angelo Salatino, Enrico Motta, Sahar Vahadati, and Jens Lehmann. Exploring large language models for scientific question answering via natural language to sparql translation. *ACM Transactions on Intelligent Systems and Technology*, 2025.

[44] Lars-Peter Meyer, Johannes Frey, Kurt Junghanns, Felix Brei, Kirill Bulert, Sabine Gründer-Fahrer, and Michael Martin. Developing a scalable benchmark for assessing large language models in knowledge graph engineering. In *SEMANTiCS 2023 (Poster, CEUR-WS Vol-3526)*, 2023.

[45] Lars-Peter Meyer et al. Assessing sparql capabilities of large language models. In *NLP4KGc @ SEMANTiCS 2024*, 2024.

[46] Nandana Mihindukulasooriya, Sanju Tiwari, Carlos F. Enguix, and Kusum Lata. Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text. In *ISWC 2023 (Springer LNCS)*, 2023. doi: 10.1007/978-3-031-47243-5_14.

[47] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP 2023*, 2023.

[48] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *ACL 2021*, 2021. doi: 10.18653/v1/2021.acl-long.416.

[49] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *EMNLP 2020*, 2020. doi: 10.18653/v1/2020.emnlp-main.154.

[50] Nathalia Nascimento, Everton Guimarães, Sai Sanjna Chintakunta, and Santhosh Anitha Boominathan. Llm4ds: Evaluating large language models for data science code generation. 2024.

[51] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *ACL 2020*, 2020. doi: 10.18653/v1/2020.acl-main.441.

[52] Guilherme Nunes, Vitor Rolla, Duarte Pereira, Vasco Alves, Andre Carreiro, and Márcia Baptista. Benchmarking table extraction: Multimodal llms vs traditional ocr. In *XLLM Workshop @ ACL 2025*, 2025.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

[53] Reham Omar, Omij Mangukiya, and Essam Mansour. Dialogue benchmark generation from knowledge graphs with cost-effective retrieval-augmented llms. In *SIGMOD 2025*, 2025. doi: 10.1145/3709681.

[54] Liu Pai, Wenyang Gao, Wenjie Dong, Lin Ai, Ziwei Gong, Songfang Huang, Li Zongsheng, Ehsan Hoque, Julia Hirschberg, and Yue Zhang. A survey on open information extraction from rule-based model to large language model. *Findings of the association for computational linguistics: EMNLP 2024*, pages 9586–9608, 2024.

[55] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. In *CoNLL 2023*, 2023.

[56] Huitong Pan, M. Adamu, Qi Zhang, Eduard Dragut, et al. Climateie: A dataset for climate science information extraction. In *ClimateNLP Workshop @ ACL 2025*, 2025. doi: 10.18653/v1/2025.climatenlp-1.6.

[57] Jeff Z Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, et al. Large language models and knowledge graphs: Opportunities and challenges. *arXiv preprint arXiv:2308.06374*, 2023.

[58] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. Bbq: A hand-built bias benchmark for question answering. In *ACL 2022 Findings*, 2022.

[59] Akshay Paruchuri, Jake Garrison, Shun Liao, John Hernandez, Jacob E. Sunshine, et al. What are the odds? language models are capable of probabilistic reasoning. In *EMNLP 2024*, 2024. doi: 10.18653/v1/2024.emnlp-main.654.

[60] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge graphs: Opportunities and challenges. *Artificial intelligence review*, page 1, 2023.

[61] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, et al. Language models as knowledge bases? In *EMNLP 2019*, 2019.

[62] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, et al. Kilt: a benchmark for knowledge intensive language tasks. In *NAACL 2021*, 2021. doi: 10.18653/v1/2021.naacl-main.200.

[63] Andrija Poleksić and Sanda Martinčić-Ipšić. Pretraining and evaluation of bert models for climate research. *Discover Applied Sciences (Springer), 2025*, 2025. doi: 10.1007/s42452-025-07740-5.

[64] Tidor-Vlad Pricope. Hardml: A benchmark for evaluating data science and machine learning knowledge and reasoning in ai. *Studia Universitatis Babeş-Bolyai Informatica, 69(2)*, 2024. doi: 10.24193/subbi.2024.2.04.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

[65] Tyler Thomas Procko and Omar Ochoa. Graph retrieval-augmented generation for large language models: A survey. In *2024 Conference on AI, science, engineering, and technology (AIxSET)*, pages 166–169. IEEE, 2024.

[66] A B M Ashikur Rahman, Saeed Anwar, Muhammad Usman, and Ajmal Mian. Defan: Definitive answer dataset for llms hallucination evaluation. 2024.

[67] Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. Lynx: An open source hallucination evaluation model. 2024.

[68] Alex Robertson, Huizhi Liang, Mahbub Gani, Rohit Kumar, and Srijith Rajamohan. Kghalubench: A knowledge graph-based hallucination benchmark for evaluating the breadth and depth of llm knowledge. In *Findings of EACL 2026*, 2026. doi: 10.18653/v1/2026.findings-eacl.206.

[69] Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, and Juba Nait Saada. Grapheval: A knowledge-graph based llm hallucination evaluation framework. In *KiL Workshop @ KDD 2024*, 2024.

[70] Tobias Schimanski, Imene Kolli, Yu Fan, Ario Saeid Vaghefi, et al. pdfqa: Diverse, challenging, and realistic question answering over pdfs. 2026.

[71] Priyanka Sen, Alham Fikri Aji, and Amir Saffari. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *COLING 2022*, 2022.

[72] Ouyang Shuyin, Huang Dong, Guo Jingwen, Sun Zeyu, Zhu Qihao, et al. Dscodebench: A realistic benchmark for data science code generation. In *AAAI 2026*, 2026. doi: 10.1609/aaai.v40i38.40540.

[73] Aarohi Srivastava and others (450 authors across 132 institutions). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TMLR 2023*, 2023.

[74] Yuan Sui, Yixiao He, Zifeng Ding, and Bryan Hooi. Can knowledge graphs make large language models more trustworthy? an empirical study over open-ended question answering. In *ACL 2025*, 2025.

[75] Sönke Tenckhoff et al. Llmstructbench: Benchmarking llms for structured data extraction. 2026.

[76] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *NAACL 2018*, 2018. doi: 10.18653/v1/N18-1074.

[77] Shiyu Tian, Yangyang Luo, Tianze Xu, Caixia Yuan, Huixing Jiang, Chen Wei, and Xiaojie Wang. Kg-adapter: Enabling knowledge graph integration in large language models through parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3813–3828, 2024.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

[78] Stefani Tsaneva, Danilo Dessì, Francesco Osborne, and Marta Sabou. Knowledge graph validation by integrating llms and human-in-the-loop. *Information Processing & Management*, 62(5):104145, 2025.

[79] Ismael Villanueva-Miranda et al. Clinbench: A standardized multi-domain framework for evaluating large language models in clinical information extraction. In *NeurIPS 2025 (Datasets and Benchmarks Track)*, 2025.

[80] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS 2019*, 2019.

[81] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS 2023*, 2023.

[82] Jason Wei, Nguyen Karina, Hyung Won Chung, et al. Measuring short-form factuality in large language models. 2024.

[83] Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Cheng Niu, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *ACL 2024*, 2024.

[84] Zhishang Xiang et al. When to use graphs in rag: A comprehensive analysis for graph retrieval-augmented generation. In *ICLR 2026*, 2026.

[85] Yunxiang Xiao et al. Graphrag-bench: A comprehensive benchmark for graph-enhanced retrieval-augmented generation. 2025.

[86] Junjie Xu et al. mmrag: A modular benchmark for retrieval-augmented generation over text, tables, and knowledge graphs. In *ISWC 2025 (Springer LNCS 16141)*, 2025. doi: 10.1007/978-3-032-09530-5_1.

[87] Dongil Yang, Minjin Kim, Sunghwan Kim, et al. Llm meets scene graph: Can large language models understand and generate scene graphs? a benchmark and empirical study. In *ACL 2025*, 2025.

[88] Jialin Yang et al. Structeval: Benchmarking llms' capabilities to generate structural outputs. *Transactions on Machine Learning Research*, 2025.

[89] Xiao Yang, Kai Sun, Hao Xin, et al. Crag – comprehensive rag benchmark. In *NeurIPS 2024*, 2024.

[90] Zhilin Yang, Peng Qi, Saizheng Zhang, et al. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP 2018*, 2018.

[91] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *ACL 2019*, 2019. doi: 10.18653/v1/P19-1472.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

[92] Qiuhai Zeng, Claire Jin, Xinyue Wang, Yuhan Zheng, and Qunhua Li. Airepr: An analyst-inspector framework for evaluating reproducibility of llms in data science. In *EMNLP 2025 Findings*, 2025. doi: 10.18653/v1/2025.findings-emnlp.539.

[93] Jie Zhang, Cezara Petrui, Kristina Nikolić, and Florian Tramèr. Realmath: A continuous benchmark for evaluating language models on research-level mathematics. 2025.

[94] Xiao Yu Cindy Zhang, Carlos R. Ferreira, Francis Rossignol, Raymond T. Ng, Wyeth Wasserman, and Jian Zhu. Casereportbench: An llm benchmark dataset for dense information extraction in clinical case reports. In *CHIL 2025 (PMLR 287)*, 2025.

[95] Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, et al. Abstractive visual understanding of multi-modal structured knowledge: A new perspective for mllm evaluation. In *ACM Multimedia 2025*, 2025. doi: 10.1145/3746027.3758158.

[96] Yuge Zhang, Qiyang Jiang, Xingyu Han, Nan Chen, Yuqing Yang, et al. Benchmarking data science agents. In *ACL 2024*, 2024. doi: 10.18653/v1/2024.acl-long.308.

[97] Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In *AAAI 2018*, 2018. doi: 10.1609/aaai.v32i1.12057.

[98] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL 2018*, 2018.

[99] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS 2023*, 2023.

[100] Dongzhuoran Zhou, Yuqicheng Zhu, Xiaxia Wang, Hongkuan Zhou, Yuan He, Jiaoyan Chen, Steffen Staab, and Evgeny Kharlamov. What breaks knowledge graph based RAG? benchmarking and empirical insights into reasoning under incomplete knowledge. In Vera Demberg, Kentaro Inui, and Lluís Marquez, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2522–2538, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.114. URL https://aclanthology.org/2026.eacl-long.114/.

[101] Yizhang Zhu, Shiyi Du, Boyan Li, Yuyu Luo, and Nan Tang. Are large language models good statisticians? In *NeurIPS 2024*, 2024.

Funded by the European Union
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY