Novelty Assessment Report

Paper: ProfBench: Multi-Domain Rubrics requiring Professional Knowledge to Answer and Judge

PDF URL: https://openreview.net/pdf?id=VwNzKPqBxk

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Evaluating progress in large language models (LLMs) is often constrained by the challenge of verifying responses, limiting assessments to tasks like mathematics, programming, and short-form question-answering. However, many real-world applications require evaluating LLMs in processing professional documents, synthesizing information, and generating comprehensive reports in response to user queries. We introduce ProfBench: a set of over 7000 response-criterion pairs as evaluated by human-experts with professional knowledge across Physics PhD, Chemistry PhD, Finance MBA and Consulting MBA. We build robust and affordable LLM-Judges to evaluate ProfBench rubrics, by mitigating self-enhancement bias and reducing the cost of evaluation by 2-3 orders of magnitude, to make it fair and accessible to the broader community. Our findings reveal that ProfBench poses significant challenges even for state-of-the-art LLMs, with top-performing models like GPT-5-high achieving only 65.9\% overall performance. Furthermore, we identify notable performance disparities between proprietary and open-weight models and provide insights into the role that extended thinking plays in addressing complex, professional-domain tasks.

Core Task Landscape

This paper addresses: **Evaluating Large Language Models on Professional Domain Tasks Requiring Expert Knowledge**

A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Healthcare and Clinical Medicine Evaluation**
- **Scientific and Technical Domain Evaluation**
- **Business and Legal Domain Evaluation**
- **Professional Task Automation and Agent Evaluation**
- **Software Engineering and Code Generation Evaluation**
- **Cross-Domain and Multi-Disciplinary Evaluation Frameworks**
- **Domain Specialization and Adaptation Methods**
- **Specialized Domain Benchmarks**
- **Expert Annotation and Human-AI Collaboration Evaluation**

Complete Taxonomy Tree

- Evaluating Large Language Models on Professional Domain Tasks Requiring Expert Knowledge Survey Taxonomy
- Healthcare and Clinical Medicine Evaluation
  - Medical Knowledge and Question Answering Assessment (2 papers)
  - [1] Large language models encode clinical knowledge (K. Singhal, 2023) View paper.)
  - [5] Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise (Qimin Yang, 2025) View paper
  - Clinical Application and Interaction Evaluation (4 papers)
  - [4] Healthbench: Evaluating large language models towards improved human health (Wei, 2025) View paper
  - [25] Evaluation of the Performance of Three Large Language Models in Clinical Decision Support: A Comparative Study Based on Actual Cases (Xueqi Wang, 2025) View paper
  - [30] Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge (Bufang Yang, 2024) View paper
  - [50] LLM-as-a-Fuzzy-Judge: Fine-Tuning Large Language Models as a Clinical Evaluation Judge with Fuzzy Logic (Zheng Weibing, 2025) View paper
  - Healthcare LLM Survey and Review Studies (3 papers)
  - [2] Large language models in healthcare and medical domain: A review (Zabir Al Nazi, 2024) View paper
  - [10] A systematic review of large language model (LLM) evaluations in clinical medicine (Sina Shool, 2025) View paper
  - [32] A comprehensive survey on evaluating large language model applications in the medical industry (Huang Yi-ning, 2024) View paper
- Scientific and Technical Domain Evaluation
  - Chemistry and Molecular Science Assessment (3 papers)
  - [7] What can large language models do in chemistry? a comprehensive benchmark on eight tasks (Guo, 2023) View paper
  - [17] Are large language models superhuman chemists? (Jablonka, 2024) View paper
  - [28] A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists (Adrian Mirza, 2025) View paper
  - Genomics and Biological Science Evaluation (1 papers)

- [33] Benchmarking large language models for genomic knowledge with GeneTuring. (Xinyi Shang, 2025) View paper
- Engineering and Cross-Scientific Evaluation (2 papers)
- [16] Engibench: A benchmark for evaluating large language models on engineering problem solving (Zhou Xi-yuan, 2025) View paper
- [47] OmniScience: A Domain-Specialized LLM for Scientific Reasoning and Discovery (Islam Md. Amirul, 2025) View paper

- Business and Legal Domain Evaluation
  - Legal Knowledge and Reasoning Assessment (2 papers)
  - [24] Lawbench: Benchmarking legal knowledge of large language models (Zhiwei Fei, 2024) View paper
  - [26] Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models (Neel Guha, 2023) View paper
  - Financial and Business Consulting Evaluation (1 papers)
  - [20] Surpassing human counterparts: A breakthrough achievement of large language models in professional tax qualification examinations in china (Lifeng Xu, 2024) View paper
- Professional Task Automation and Agent Evaluation
  - General Task Automation and Workflow Evaluation (1 papers)
  - [3] Taskbench: Benchmarking large language models for task automation (Dongsheng Li, 2024) View paper
  - Domain-Specific Professional Workflow Evaluation (3 papers)
  - [22] Can large language models replace human experts? Effectiveness and limitations in building energy retrofit challenges assessment (Linyan Chen, 2025) View paper
  - [34] Legalagentbench: Evaluating llm agents in legal domain (Haitao Li, 2025) View paper
  - [41] CRMArena: Understanding the Capacity of LLM Agents to Perform Professional CRM Tasks in Realistic Environments (Kung-Hsiang Huang, 2024) View paper
- Software Engineering and Code Generation Evaluation
  - Code Generation and Database Interface Assessment (1 papers)
  - [35] Can LLM Already Serve as A Database Interface? A BIg Bench for Large-Scale Database Grounded Text-to-SQLs (Li Jinyang, 2023) View paper
  - Software Engineering Communication and Collaboration Evaluation (1 papers)
  - [48] Humanevalcomm: Benchmarking the communication competence of code generation for llms and llm agents (Jie Jw Wu, 2025) View paper
  - Software Engineering Survey and Review Studies (2 papers)
  - [37] From llms to llm-based agents for software engineering: A survey of current, challenges and future (Jin Haolin, 2024) View paper
  - [38] Assessing and Advancing Benchmarks for Evaluating Large Language Models in Software Engineering Tasks (Hu Xing, 2025) View paper
- Cross-Domain and Multi-Disciplinary Evaluation Frameworks
  - Multi-Domain Professional Knowledge Benchmarks ★ (3 papers)
  - [0] ProfBench: Multi-Domain Rubrics requiring Professional Knowledge to Answer and Judge (Anon et al., 2026) View paper
  - [18] ExpertLongBench: Benchmarking Language Models on Expert-Level Long-Form Generation Tasks with Structured Checklists (Ruan Jie, 2025) View paper
  - [44] Supergpqa: Scaling llm evaluation across 285 graduate disciplines (P. Team, 2025) View paper
  - Cross-Lingual and Multi-Modal Professional Evaluation (2 papers)
  - [13] A comparative study on large language models' accuracy in cross-lingual professional terminology processing: An evaluation across multiple domains (H Zhang, 2024) View paper
  - [45] MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? (Zhang Yi Fan, 2024) View paper
  - Evaluation Methodology and Meta-Evaluation Frameworks (3 papers)
  - [9] Benchmarking foundation models with language-model-as-an-examiner (Bai, 2023) View paper
  - [11] Evaluation Workflows for Large Language Models (LLMs) that Integrate Domain Expertise for Complex Knowledge Tasks (Annalisa Szymanski, 2025) View paper
  - [43] Judgebench: A benchmark for evaluating llm-based judges (Tan, 2024) View paper
  - General LLM Evaluation Survey Studies (1 papers)
  - [8] A survey on evaluation of large language models (Yu-Peng Chang, 2024) View paper
- Domain Specialization and Adaptation Methods
  - Domain-Specific Training and Fine-Tuning Approaches (3 papers)
  - [12] Adapting large language models to domains via reading comprehension (Cheng, 2023) View paper
  - [39] Aligning language models to professional domains using preference training (HarÃ°arson, 2024) View paper
  - [46] Fine-tuning and utilization methods of domain-specific llms (Cheon-Su Jeong, 2024) View paper
  - Knowledge Integration and Retrieval-Augmented Methods (5 papers)
  - [23] Way to specialist: Closing loop between specialized llm and evolving domain knowledge graph (Yutong Zhang, 2025) View paper
  - [29] From human experts to machines: An LLM supported approach to ontology and knowledge graph construction (Kommineni, 2024) View paper
  - [31] EXPERTSTEER: Intervening in LLMs through Expert Knowledge (Wang Wei-xuan, 2025) View paper
  - [40] Injecting domain-specific knowledge into large language models: a comprehensive survey (Zirui Song, 2025) View paper
  - [49] Enhancing large language models through external domain knowledge (L. Welz, 2024) View paper
  - Domain-Specialized Architecture and Knowledge Partitioning (1 papers)
  - [19] Large language models with knowledge domain partitioning for specialized domain knowledge concentration (Xijun Gong, 2024) View paper
  - Domain Specialization Survey and Review Studies (2 papers)
  - [14] Domain specialization as the key to make large language models disruptive: A comprehensive survey (Chen Ling, 2025) View paper
  - [27] Survey of specialized large language model (Yang ChenGhan, 2025) View paper
- Specialized Domain Benchmarks
  - Humanities and Social Science Domain Benchmarks (1 papers)

- [42] Large language models' expert-level global history knowledge benchmark (HiST-LLM) (Majid Benam, 2024) View paper
- Industry-Specific Professional Domain Benchmarks (2 papers)
- [15] ElecBench: A large language model benchmark in electric power domain (Sai Zhang, 2025) View paper
- [21] A multi-dimensional performance evaluation of large language models in dental implantology: comparison of ChatGPT, DeepSeek, Grok, Gemini and Qwen across â (X Wu, 2025) View paper

• Expert Annotation and Human-AI Collaboration Evaluation (2 papers)
- [6] Evaluating Large Language Models as Expert Annotators (Chen-Wei Lin, 2025) View paper
- [36] Idea Evaluation for Solutions to Specialized Problems: Leveraging the Potential of Crowds and Large Language Models (Henner Gimpel, 2025) View paper

Narrative

Core task: evaluating large language models on professional domain tasks requiring expert knowledge. The field has organized itself into several major branches that reflect both the diversity of professional domains and the methodological challenges of rigorous evaluation. Healthcare and Clinical Medicine Evaluation (e.g., HealthBench[4], Clinical LLM Review[10]) focuses on diagnostic reasoning and medical decision-making, while Scientific and Technical Domain Evaluation spans chemistry, engineering, and other STEM fields where specialized terminology and problem-solving are paramount. Business and Legal Domain Evaluation (e.g., LawBench[24], LegalBench[26]) addresses regulatory compliance and contract analysis, and Software Engineering and Code Generation Evaluation examines programming tasks. Cross-Domain and Multi-Disciplinary Evaluation Frameworks aim to assess models across multiple professional areas simultaneously, complemented by branches on Domain Specialization and Adaptation Methods that explore fine-tuning and knowledge injection strategies, Specialized Domain Benchmarks that provide targeted test suites, and Expert Annotation and Human-AI Collaboration Evaluation that investigates how domain experts interact with and validate model outputs.

A central tension runs through these branches: whether to build narrow, deeply specialized benchmarks for individual professions or to create broader frameworks that capture transferable reasoning skills across domains. Works like TaskBench[3] and LLM Evaluation Survey[8] emphasize general-purpose evaluation paradigms, while others such as Chemistry Benchmark[7] and ElecBench[15] drill into discipline-specific nuances. ProfBench[0] sits within the Cross-Domain and Multi-Disciplinary Evaluation Frameworks branch, positioning itself alongside ExpertLongBench[18] and SuperGPQA[44] as a multi-domain professional knowledge benchmark. Compared to ExpertLongBench[18], which emphasizes long-context reasoning across expert fields, ProfBench[0] appears to prioritize breadth of professional coverage and the integration of expert-level task diversity, reflecting ongoing debates about whether comprehensive multi-domain assessments can meaningfully capture the depth that single-domain benchmarks provide.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. ExpertLongBench: Benchmarking Language Models on Expert-Level Long-Form Generation Tasks with Structured Checklists

Authors: Ruan Jie, Nair, Inderjeet, Jie Ruan, Cao Shuyang, et al. (29 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

This paper introduces ExpertLongBench, an expert-level benchmark containing 11 tasks from 9 domains that reflect realistic expert workflows and applications. Beyond question answering, the application-driven tasks in ExpertLongBench demand long-form outputs that can exceed 5,000 tokens and strict adherence to domain-specific requirements. Notably, each task in ExpertLongBench includes a rubric, designed or validated by domain experts, to specify task requirements and guide output evaluation. Fur...

Relationship Analysis

Both papers belong to the Multi-Domain Professional Knowledge Benchmarks category, creating comprehensive evaluation frameworks with expert-validated rubrics across multiple professional domains. They overlap in their focus on assessing LLMs using structured rubrics for graduate/professional-level tasks and employing LLM-judges for evaluation. However, ProfBench emphasizes rubric-based evaluation across 4 domains (Physics PhD, Chemistry PhD, Finance MBA, Consulting MBA) with 7,347 response-criterion pairs focused on binary criterion fulfillment, while ExpertLongBench covers 11 tasks across 9 domains with 1,050 samples, specifically targeting long-form generation tasks (outputs exceeding 5,000 tokens) and introducing a checklist-based evaluation framework (CLEAR) that extracts and compares itemized content rather than binary judgments.

---

2. Supergpqa: Scaling llm evaluation across 285 graduate disciplines

Authors: P. Team, M-A-P Team, Xinrun Du, Du Xinrun, Yao Yifan, et al. (230 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large language models (LLMs) have demonstrated remarkable proficiency in mainstream academic disciplines such as mathematics, physics, and computer science. However, human knowledge encompasses over 200 specialized disciplines, far exceeding the scope of existing benchmarks. The capabilities of LLMs in many of these specialized fields-particularly in light industry, agriculture, and service-oriented disciplines-remain inadequately evaluated. To address this gap, we present SuperGPQA, a comprehen...

Relationship Analysis

Both papers belong to the Multi-Domain Professional Knowledge Benchmarks category, creating comprehensive evaluation frameworks that assess LLM capabilities across multiple professional domains using expert-validated criteria and graduate-level knowledge tasks. They overlap in their focus on evaluating LLMs on challenging, expert-level questions spanning diverse professional fields (ProfBench covers 4 domains with 7000+ response-criterion pairs; SuperGPQA covers 285 graduate disciplines with 26,529 questions). The key difference is that ProfBench emphasizes rubric-based evaluation with detailed human-written criteria for open-ended responses and LLM-judge validation, while SuperGPQA focuses on multiple-choice question format across a much broader range of disciplines (285 vs 4) with human-LLM collaborative filtering to ensure question difficulty and discrimination.

Contributions Analysis

Overall novelty summary. ProfBench introduces a multi-domain professional knowledge benchmark spanning Physics PhD, Chemistry PhD, Finance MBA, and Consulting MBA tasks, with over 7000 response-criterion pairs validated by human experts. The paper resides in the Multi-Domain Professional Knowledge Benchmarks leaf, which contains only three papers including ProfBench itself, ExpertLongBench, and SuperGPQA. This sparse leaf within the Cross-Domain and Multi-Disciplinary Evaluation Frameworks branch suggests the work addresses a relatively underexplored research direction—comprehensive multi-domain professional evaluation remains less crowded than single-domain benchmarks found in Healthcare or Scientific evaluation branches.

The taxonomy reveals substantial activity in adjacent single-domain evaluation branches: Healthcare and Clinical Medicine Evaluation contains nine papers across three sub-categories, while Scientific and Technical Domain Evaluation spans six papers covering chemistry, genomics, and engineering. Business and Legal Domain Evaluation includes three papers focused on legal reasoning and financial analysis. ProfBench's multi-domain approach contrasts with these specialized branches by attempting to capture transferable reasoning across professional fields rather than drilling into discipline-specific nuances. The Cross-Domain parent branch also houses evaluation

methodology frameworks and cross-lingual assessments, indicating growing interest in general-purpose evaluation paradigms beyond domain-specific test suites.

Among 29 candidates examined through limited semantic search, the contribution-level analysis reveals varied novelty signals. The core ProfBench benchmark contribution examined 9 candidates with no clear refutations, suggesting the specific combination of expert-created rubrics across these four professional domains represents relatively novel ground. Performance measurement of 40+ models examined 10 candidates without refutation, indicating this systematic comparison may offer new empirical insights. However, methods to reduce LLM-Judge bias examined 10 candidates and found 1 refutable match, suggesting prior work exists on bias mitigation and cost reduction in LLM-based evaluation, though the specific techniques applied to professional domain rubrics may still contribute incremental value.

Based on this limited search scope covering 29 semantically similar papers, ProfBench appears to occupy a moderately novel position within a sparse research direction. The multi-domain professional benchmark itself shows stronger novelty signals than the LLM-Judge methodology components. The analysis does not cover exhaustive literature on evaluation frameworks or domain-specific benchmarks outside the top-K semantic matches, so definitive claims about absolute novelty remain constrained by search boundaries.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: ProfBench benchmark with expert-created rubrics across multiple professional domains

Description: The authors present ProfBench, a new benchmark containing over 7000 response-criterion pairs evaluated by human experts across four professional domains: Physics PhD, Chemistry PhD, Finance MBA, and Consulting MBA. This benchmark enables evaluation of LLMs on challenging, real-world professional tasks requiring domain expertise.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Expert evaluation of large language models for clinical dialogue summarization

URL: View paper

Brief Assessment

Clinical Dialogue Evaluation[73] focuses on evaluating LLMs for clinical dialogue summarization using expert clinician evaluations, not on creating multi-domain professional benchmarks with expert rubrics for general professional tasks.

---

2. Evaluation of Reliability Criteria for News Publishers with Large Language Models

URL: View paper

Brief Assessment

News Publisher Reliability[77] focuses on evaluating news article reliability using expert-designed criteria for journalism quality, not on creating a multi-domain professional benchmark for evaluating LLM capabilities across diverse expert fields like physics, chemistry, finance, and consulting.

---

3. Rubrics as rewards: Reinforcement learning beyond verifiable domains

URL: View paper

Brief Assessment

Rubrics as Rewards[79] focuses on using rubrics as reward signals for on-policy reinforcement learning training, not on creating a benchmark with expert rubrics across professional domains. The candidate evaluates on existing benchmarks (HealthBench, GPQA) rather than introducing a new multi-domain professional benchmark.

---

4. A Scalable Framework for Evaluating Health Language Models

URL: View paper

Brief Assessment

Health Evaluation Framework[76] focuses specifically on evaluating health language models using rubrics for health-related queries, not professional domains like Physics PhD, Chemistry PhD, Finance MBA, and Consulting MBA that ProfBench covers.

---

5. Scalable evaluation framework for retrieval augmented generation in tobacco research using large Language models

URL: View paper

Brief Assessment

Tobacco Research RAG[74] focuses on evaluating LLMs specifically for tobacco research document retrieval using automated metrics and expert validation, not on creating expert rubrics across diverse professional domains for general LLM evaluation.

---

6. Aecbench: A hierarchical benchmark for knowledge evaluation of large language models in the aec field

URL: View paper

Brief Assessment

AECBench[72] focuses exclusively on the architecture, engineering, and construction (AEC) domain with a hierarchical cognitive framework, while ProfBench covers multiple professional domains (Physics PhD, Chemistry PhD, Finance MBA, Consulting MBA). The domain focus and evaluation scope differ fundamentally.

---

7. Towards a personal health large language model

URL: View paper

Brief Assessment

Personal Health LLM[80] focuses on personal health domains (sleep and fitness) with wearable sensor data, not professional domains like Physics PhD, Chemistry PhD, Finance MBA, and Consulting MBA that ProfBench covers.

---

8. Ucfe: A user-centric financial expertise benchmark for large language models

URL: View paper

Brief Assessment

UCFE[75] focuses on user-centric financial expertise evaluation with dynamic multi-round dialogues and user preference alignment, rather than expert-created rubrics across multiple professional domains like physics, chemistry, finance, and consulting.

---

9. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents

URL: View paper

Brief Assessment

ResearchRubrics[71] focuses on evaluating deep research agents across diverse domains with expert-written rubrics, but targets different task types (open-ended web exploration and synthesis) compared to ProfBench's professional domain tasks requiring PhD/MBA-level knowledge. The benchmarks serve complementary purposes in the evaluation landscape.

Contribution 2: Performance measurement of over 40 models as report-generators and LLM-Judges

Description: The authors evaluate more than 40 language models both as generators of professional reports and as judges that assess whether responses meet expert-defined criteria. They analyze trends across open/closed-source models, reasoning/instruct models, and model sizes.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Generative AI and misinformation: a scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation

URL: View paper

Brief Assessment

Misinformation Scoping Review[64] focuses on evaluating LLMs for misinformation detection, generation, mitigation, and impact assessment—not on professional report generation or rubric-based judging across diverse domains like ProfBench.

---

2. A survey of textual cyber abuse detection using cutting-edge language models and large language models

URL: View paper

Brief Assessment

Cyber Abuse Detection[67] focuses on evaluating language models for detecting various forms of textual cyber abuse (hate speech, cyberbullying, trolling, etc.) rather than evaluating models as both report generators and judges of response quality in professional domains. The candidate's evaluation framework is fundamentally different from the original paper's dual-role assessment methodology.

---

3. From generation to judgment: Opportunities and challenges of llm-as-a-judge

URL: View paper

Brief Assessment

Generation to Judgment[62] focuses on evaluating LLMs as judges for assessing response quality in general contexts, not on measuring performance of models as both report-generators and judges across professional domains requiring expert knowledge (physics PhD, chemistry PhD, finance MBA, consulting MBA) as in the original paper.

---

4. A survey on the use of large language models (llms) in fake news

URL: View paper

Brief Assessment

Fake News Survey[63] is a review article focused on fake news and fake profile detection using LLMs, not on evaluating models as report generators and judges across professional domains with rubric-based assessment.

---

5. Eduquick: A dataset toward evaluating summarization of informal educational content for social media

URL: View paper

Brief Assessment

EduQuick[68] focuses on evaluating LLMs for summarizing educational content for TikTok videos, not on comprehensive multi-domain professional report generation or systematic LLM-judge evaluation across diverse criteria types.

---

6. LLMs for Customized Marketing Content Generation and Evaluation at Scale

URL: View paper

Brief Assessment

Marketing Content Generation[69] focuses on evaluating LLMs for marketing ad copy generation and quality assessment in e-commerce, not on evaluating models as professional report generators across diverse academic/business domains with expert-defined rubrics.

---

7. The Dual Threat of Large Language Models: Addressing Plagiarism and Deepfake Generation

URL: View paper

Brief Assessment

Plagiarism and Deepfakes[70] focuses on plagiarism detection and deepfake text generation/detection, not on evaluating language models as report generators or judges of response quality. The candidate's evaluation tasks involve plagiarism classification and deepfake detection, which are fundamentally different from the original paper's rubric-based professional report generation and judgment framework.

---

8. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation

URL: View paper

Brief Assessment

Scientific Discovery Survey[61] is a broad survey covering AI applications across the scientific research cycle (search, experimentation, content generation, evaluation). It does not present empirical evaluations of 40+ models as both generators and judges with specific performance metrics, rubrics, or bias analysis as described in the original contribution.

---

9. Automated test creation using large language models: A practical application

URL: View paper

Brief Assessment

Automated Test Creation[65] focuses on using LLMs for educational test generation and evaluation in a specific application context, not on comprehensive benchmarking of 40+ models as both report generators and judges across professional domains.

---

10. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation

URL: View paper

Brief Assessment

Dual Role Disinformation[66] evaluates LLMs for both generating and detecting disinformation content, not for generating professional reports and judging report quality against expert-defined criteria as in the original paper.

Contribution 3: Methods to reduce LLM-Judge bias and evaluation cost

Description: The authors develop techniques to mitigate self-enhancement bias in LLM-Judges and reduce evaluation costs by 2-3 orders of magnitude. Their approach achieves no more than 1% bias across three models from different providers while costing only $12 using the o3 model.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Inadequacies of large language model benchmarks in the era of generative artificial intelligence

URL: View paper

Brief Assessment

Benchmark Inadequacies[52] focuses on critiquing existing LLM benchmarks through a unified evaluation framework, identifying systemic inadequacies in benchmark design. It does not present methods to reduce LLM-Judge bias or evaluation costs, which are the core technical contributions of the original paper.

---

2. Length-controlled alpacaeval: A simple way to debias automatic evaluators

URL: View paper

Brief Assessment

Length-controlled AlpacaEval[56] addresses length bias in automated evaluations using regression-based debiasing, while the original paper tackles self-enhancement bias and develops cost-reduction methods for rubric-based professional benchmarks. These are distinct technical approaches to different bias problems in different evaluation contexts.

---

3. Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena

URL: View paper

Brief Assessment

Open-LLM-Leaderboard[57] addresses selection bias in multiple-choice questions by transitioning to open-style questions, not LLM-Judge bias in rubric-based evaluation systems. The candidate focuses on eliminating answer choice biases (A/B/C/D preferences) rather than self-enhancement bias in judge models or cost reduction in evaluation frameworks.

---

4. Bias and fairness in large language models: A survey

URL: View paper

Brief Assessment

Bias and Fairness[53] is a survey on social bias and fairness in LLMs, focusing on harmful social biases in model outputs. It does not address LLM-Judge evaluation systems or methods to reduce self-enhancement bias in automated evaluation contexts.

---

5. Pre: A peer review based large language model evaluator

URL: View paper

Brief Assessment

Peer Review Evaluator[59] addresses bias through a peer-review mechanism using multiple LLM reviewers, while the original paper focuses on self-enhancement bias mitigation and cost reduction through specific technical methods (achieving $12 cost with o3 model). The candidate's approach is fundamentally different in methodology.

---

6. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias

URL: View paper

Brief Assessment

Training Data Bias[58] focuses on reducing bias in LLM-generated training data for text classification tasks, not on reducing bias in LLM-Judge evaluation systems. The candidate addresses attribute bias in synthetic datasets and cost efficiency of data generation, which is orthogonal to the original paper's contribution on mitigating self-enhancement bias in LLM-Judges for rubric-based evaluation.

---

7. Judgelm: Fine-tuned large language models are scalable judges

URL: View paper

Prior Art Analysis

JudgeLM[51] demonstrates prior work on reducing bias in LLM judges through systematic techniques. The candidate paper explicitly addresses key biases (position bias, knowledge bias, and format bias) and proposes specific technical solutions including swap augmentation, reference support, and other techniques to mitigate these biases. This shows that methods for reducing LLM-judge bias were developed and published before the original paper's submission. While the original paper claims to develop techniques to mitigate self-enhancement bias specifically and reduce costs by 2-3 orders of magnitude, JudgeLM[51] establishes the foundational work on identifying and addressing multiple types of bias in LLM judges, which challenges the novelty of the bias reduction approach.

Evidence

Evidence 1 - **Rationale**: Both papers address the problem of bias in LLM judges. JudgeLM[51] identifies and addresses multiple types of bias (position, knowledge, format) with specific techniques, establishing prior work on bias reduction methods in LLM judges before the original paper's submission. - **Original**: we propose methods to reduce the bias of llm-judges in favoring responses from specific providers as well as the cost of running the benchmark, in order to improve its accessibility. - **Candidate**: we then analyze the key biases in fine-tuning llm as a judge and consider them as position bias, knowledge bias, and format bias. to address these issues, judgelm introduces a bag of techniques including swap augmentation, reference support, and

---

8. Bias testing and mitigation in llm-based code generation

URL: View paper

Brief Assessment

Code Generation Bias[55] focuses on bias in LLM-generated code (social biases like gender/race in software) and mitigation through prompt engineering, not on reducing bias in LLM-as-judge evaluation systems or evaluation costs for benchmarking.

---

9. Split and merge: Aligning position biases in LLM-based evaluators

URL: View paper

Brief Assessment

Split and Merge[54] addresses position bias in LLM evaluators through answer segmentation and alignment, not self-enhancement bias. The candidate focuses on pairwise comparison consistency rather than reducing evaluation costs by orders of magnitude as claimed in the original paper.

---

10. Application of unified health large language model evaluation framework to In-Basket message replies: bridging qualitative and quantitative assessments

URL: View paper

Brief Assessment

In-Basket Evaluation[60] focuses on healthcare-specific LLM evaluation using a unified framework that bridges qualitative and quantitative assessments. It does not address LLM-Judge bias mitigation or cost reduction techniques for evaluation systems, which are the core technical contributions of the original paper.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] ProfBench: Multi-Domain Rubrics requiring Professional Knowledge to Answer and Judge View paper
- [1] Large language models encode clinical knowledge View paper.)
- [2] Large language models in healthcare and medical domain: A review View paper
- [3] Taskbench: Benchmarking large language models for task automation View paper
- [4] Healthbench: Evaluating large language models towards improved human health View paper
- [5] Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise View paper
- [6] Evaluating Large Language Models as Expert Annotators View paper
- [7] What can large language models do in chemistry? a comprehensive benchmark on eight tasks View paper
- [8] A survey on evaluation of large language models View paper
- [9] Benchmarking foundation models with language-model-as-an-examiner View paper
- [10] A systematic review of large language model (LLM) evaluations in clinical medicine View paper
- [11] Evaluation Workflows for Large Language Models (LLMs) that Integrate Domain Expertise for Complex Knowledge Tasks View paper
- [12] Adapting large language models to domains via reading comprehension View paper
- [13] A comparative study on large language models' accuracy in cross-lingual professional terminology processing: An evaluation across multiple domains View paper
- [14] Domain specialization as the key to make large language models disruptive: A comprehensive survey View paper
- [15] ElecBench: A large language model benchmark in electric power domain View paper
- [16] Engibench: A benchmark for evaluating large language models on engineering problem solving View paper
- [17] Are large language models superhuman chemists? View paper
- [18] ExpertLongBench: Benchmarking Language Models on Expert-Level Long-Form Generation Tasks with Structured Checklists View paper
- [19] Large language models with knowledge domain partitioning for specialized domain knowledge concentration View paper
- [20] Surpassing human counterparts: A breakthrough achievement of large language models in professional tax qualification examinations in china View paper
- [21] A multi-dimensional performance evaluation of large language models in dental implantology: comparison of ChatGPT, DeepSeek, Grok, Gemini and Qwen across â¦ View paper
- [22] Can large language models replace human experts? Effectiveness and limitations in building energy retrofit challenges assessment View paper
- [23] Way to specialist: Closing loop between specialized llm and evolving domain knowledge graph View paper
- [24] Lawbench: Benchmarking legal knowledge of large language models View paper
- [25] Evaluation of the Performance of Three Large Language Models in Clinical Decision Support: A Comparative Study Based on Actual Cases View paper
- [26] Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models View paper
- [27] Survey of specialized large language model View paper
- [28] A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists View paper
- [29] From human experts to machines: An LLM supported approach to ontology and knowledge graph construction View paper
- [30] Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge View paper
- [31] EXPERTSTEER: Intervening in LLMs through Expert Knowledge View paper
- [32] A comprehensive survey on evaluating large language model applications in the medical industry View paper
- [33] Benchmarking large language models for genomic knowledge with GeneTuring. View paper
- [34] Legalagentbench: Evaluating llm agents in legal domain View paper
- [35] Can LLM Already Serve as A Database Interface? A BIg Bench for Large-Scale Database Grounded Text-to-SQLs View paper
- [36] Idea Evaluation for Solutions to Specialized Problems: Leveraging the Potential of Crowds and Large Language Models View paper
- [37] From llms to llm-based agents for software engineering: A survey of current, challenges and future View paper
- [38] Assessing and Advancing Benchmarks for Evaluating Large Language Models in Software Engineering Tasks View paper
- [39] Aligning language models to professional domains using preference training View paper
- [40] Injecting domain-specific knowledge into large language models: a comprehensive survey View paper
- [41] CRMArena: Understanding the Capacity of LLM Agents to Perform Professional CRM Tasks in Realistic Environments View paper
- [42] Large language models' expert-level global history knowledge benchmark (HiST-LLM) View paper
- [43] Judgebench: A benchmark for evaluating llm-based judges View paper
- [44] Supergpqa: Scaling llm evaluation across 285 graduate disciplines View paper
- [45] MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? View paper

- [46] Fine-tuning and utilization methods of domain-specific llms View paper
- [47] OmniScience: A Domain-Specialized LLM for Scientific Reasoning and Discovery View paper
- [48] Humanevalcomm: Benchmarking the communication competence of code generation for llms and llm agents View paper
- [49] Enhancing large language models through external domain knowledge View paper
- [50] LLM-as-a-Fuzzy-Judge: Fine-Tuning Large Language Models as a Clinical Evaluation Judge with Fuzzy Logic View paper
- [51] Judgelm: Fine-tuned large language models are scalable judges View paper
- [52] Inadequacies of large language model benchmarks in the era of generative artificial intelligence View paper
- [53] Bias and fairness in large language models: A survey View paper
- [54] Split and merge: Aligning position biases in LLM-based evaluators View paper
- [55] Bias testing and mitigation in llm-based code generation View paper
- [56] Length-controlled alpacaeval: A simple way to debias automatic evaluators View paper
- [57] Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena View paper
- [58] Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias View paper
- [59] Pre: A peer review based large language model evaluator View paper
- [60] Application of unified health large language model evaluation framework to In-Basket message replies: bridging qualitative and quantitative assessments View paper
- [61] Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation View paper
- [62] From generation to judgment: Opportunities and challenges of llm-as-a-judge View paper
- [63] A survey on the use of large language models (llms) in fake news View paper
- [64] Generative AI and misinformation: a scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation View paper
- [65] Automated test creation using large language models: A practical application View paper
- [66] Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation View paper
- [67] A survey of textual cyber abuse detection using cutting-edge language models and large language models View paper
- [68] Eduquick: A dataset toward evaluating summarization of informal educational content for social media View paper
- [69] LLMs for Customized Marketing Content Generation and Evaluation at Scale View paper
- [70] The Dual Threat of Large Language Models: Addressing Plagiarism and Deepfake Generation View paper
- [71] Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents View paper
- [72] Aecbench: A hierarchical benchmark for knowledge evaluation of large language models in the aec field View paper
- [73] Expert evaluation of large language models for clinical dialogue summarization View paper
- [74] Scalable evaluation framework for retrieval augmented generation in tobacco research using large Language models View paper
- [75] Ucfe: A user-centric financial expertise benchmark for large language models View paper
- [76] A Scalable Framework for Evaluating Health Language Models View paper
- [77] Evaluation of Reliability Criteria for News Publishers with Large Language Models View paper
- [78] One for All: A General Framework of LLMs-based Multi-Criteria Decision Making on Human Expert Level View paper
- [79] Rubrics as rewards: Reinforcement learning beyond verifiable domains View paper
- [80] Towards a personal health large language model View paper