Task-Oriented Suitability Analysis of Large Language Models as Research Collaborators

Abstract

Large language models (LLMs) are increasingly used across multiple stages of scientific and engineering research, supporting tasks such as background understanding, reasoning, and experimental implementation. Despite this growing adoption, there is limited guidance on how to select an appropriate LLM for a specific research task. In practice, model selection often relies on general-purpose benchmark rankings or individual experience, which may not adequately reflect the diverse requirements of research activities. This study formulates the use of LLMs in research as a task-oriented decision problem and presents a descriptive analytical approach for examining model utilization characteristics. Using controlled and repeated experiments with public benchmarks and fixed prompts, we reinterpret representative benchmark results from the perspective of research task requirements. Reasoning-oriented benchmarks, including MMLU and GSM8K, are analyzed in relation to theory- and analysis-focused research tasks, while code generation benchmarks such as HumanEval are examined in the context of experimentation and simulation-oriented tasks. Based on an explicit task–benchmark mapping, we compare performance distributions and response characteristics across multiple LLM families under identical evaluation conditions. Rather than producing a single performance ranking, the analysis organizes recurring response tendencies and limitation patterns observed at the task level. The results indicate that LLM behavior varies substantially across research task types and that overall benchmark scores alone do not sufficiently characterize model suitability in research settings. By framing LLM usage as a task-conditioned analytical problem, this study provides a reproducible and interpretable perspective for understanding how different models behave across research contexts, offering descriptive reference criteria that may support researchers in making more informed model selection decisions.

1 Introduction

1.1 Problem Statement and Motivation

Large language models (LLMs) are now used across a broad range of activities in scientific and engineering research, including literature exploration, organization of theoretical background, logical and mathematical reasoning, experimental code development, and result summarization [1]–[3]. As their capabilities have expanded, LLMs have become integrated into research workflows not only as auxiliary tools but as systems that provide continuous support for multiple research tasks [4], [5]. In particular, the adoption of LLMs during the early and exploratory stages of research has increased, raising practical questions about how different models behave when applied to different types of research tasks.

Despite this widespread use, there is limited consensus on how to select an appropriate LLM for a given research task. In practice, researchers often refer to public leaderboard rankings, a small set of benchmark scores, or prior personal experience when choosing a model [6]–[8]. While these sources offer coarse indicators of general model performance, they do not directly capture the heterogeneous requirements of concrete research activities. Research tasks differ substantially in their primary demands, ranging from theory-driven analysis and multi-step reasoning to specification-based code

Submitted to 1st 2026AI Co-Scientist Challenge Korea. Do not distribute.

implementation, simulation, and tasks that require strict adherence to predefined formats or protocols [9]–[12].

As a result, mismatches between model capabilities and task requirements are frequently observed in practical research settings. Models that exhibit relatively stable behavior in conceptual organization or reasoning-oriented tasks may produce recurrent errors in code generation or fail to comply with structural constraints. Conversely, models that perform effectively in implementation-oriented tasks may generate less coherent or less useful outputs when applied to complex analytical reasoning or problem formulation. Such discrepancies can lead to repeated trial-and-error during research execution, increasing time and effort without providing clear guidance on model selection.

These observations suggest that the challenge in LLM utilization is not solely a matter of identifying a universally superior model, but rather of understanding how model response characteristics align with different research task types. From this perspective, LLM selection can be viewed as a task-conditioned decision problem, in which suitability depends on the interaction between task requirements and model behavior, rather than on a single aggregate performance measure.

Accordingly, there is a need for analytical approaches that examine LLM behavior under reproducible conditions and interpret benchmark results in relation to research task characteristics. Analyses based on publicly available benchmarks, when conducted with controlled input settings, can provide descriptive evidence about how different models respond across task types. Such task-oriented interpretations may help reduce uncertainty in model selection and support more informed use of LLMs as research collaborators, particularly as research workflows increasingly involve sustained and repeated interaction with these models over extended periods [13], [14].

1.2 Contributions

This study examines the use of large language models (LLMs) as research collaborators from a task-oriented analytical perspective, focusing on how model response characteristics vary across different types of research tasks. Rather than treating LLM evaluation as a single-score performance comparison, the study analyzes LLM utilization as a task-conditioned research analysis problem grounded in observable response behavior. The main contributions of this work are summarized as follows.

First, this study moves beyond evaluation perspectives centered on aggregate performance rankings and analyzes LLM selection in relation to research task types rather than absolute scores [6], [8]. By distinguishing among theory- and analysis-oriented tasks, experimentation- and simulation-oriented tasks, and tasks that require strict adherence to research protocols or output formats, the study highlights that different research contexts impose distinct capability demands on LLMs [7], [9]. This formulation emphasizes that observed model performance tendencies and limitations vary across task settings and that LLM behavior cannot be adequately characterized by a single overall ranking.

Second, this study introduces a task–benchmark mapping that reinterprets representative public benchmarks from a research task perspective. Under the assumption that reasoning-focused, code-generation-focused, and instruction-following benchmarks reflect different core capabilities required at different stages of research, the analysis explicitly links benchmark outcomes to corresponding research activities [10], [11]. This mapping provides an analytical basis for interpreting which aspects of research task execution are emphasized by existing benchmark scores, without treating those scores as comprehensive indicators of research suitability.

Third, the study presents an agent-based analysis methodology that systematically organizes observed performance distributions and recurring response patterns across multiple LLMs under controlled conditions. The agent evaluates multiple models using identical benchmark inputs, fixed prompts, and consistent evaluation settings, and aggregates task-level performance indicators together with recurrent failure patterns [15], [16]. These structured summaries enable descriptive comparison of model response stability and limitation tendencies across different research task types.

Fourth, the study adopts a reproducible experimental design based on publicly available benchmark datasets and fixed execution settings. All experiments are conducted under identical input conditions, allowing the analysis to be repeated and independently examined by third parties. This design supports objective comparison across models and ensures that the observed task-level response characteristics are not artifacts of prompt variation or implementation-specific factors.

Finally, this work explicitly constrains the role of AI to that of a controlled analytical support agent rather than a decision maker or evaluation authority. The agent does not intervene in model response generation or judgment, but instead operates as a post hoc analysis layer that organizes experimental outputs in a comparable and structured manner [17]. Through this design, the study illustrates how LLM-based automation can support researcher analysis while preserving human judgment in research decision-making.

2 Background and Terminology

2.1 Large Language Models as Research Collaborators

Large Language Models (LLMs) are general-purpose models trained on large-scale text corpora and are widely applied to natural language understanding and generation, code synthesis, and reasoning tasks [18], [19], [20], [21], [22]. Recently, their role has expanded beyond standalone question-answering tools, and LLMs are increasingly used throughout the research workflow to support analysis, content generation, and review. In this study, a research collaborator is defined as a tool that assists researchers by improving efficiency and productivity without replacing human decision-making. Although prior studies have examined the use of LLMs for individual functions such as literature search, code generation, and result summarization [1], [2], these capabilities are typically combined within concrete research tasks. Consequently, evaluating LLMs solely based on isolated abilities provides limited insight into their effectiveness in practical research settings.

2.2 Limitations of LLM Evaluation and Benchmarks

Public benchmarks have been widely used to evaluate LLMs in terms of knowledge understanding, reasoning, code generation, and instruction following [23], [24], [25], [26], and they serve as important tools for model comparison and progress tracking [6]. However, most benchmarks assess individual capability dimensions independently and do not capture the integrated and context-dependent requirements of real-world research tasks [8], [27]. Moreover, benchmark outcomes are often summarized as single scores or rankings, which may misleadingly suggest universal model superiority. Such interpretations overlook task diversity and provide limited guidance for selecting LLMs tailored to specific research needs.

2.3 Task-Oriented Perspective and the Concept of Suitability

A research task is defined as a goal-oriented activity performed at a particular stage of the research process, such as theoretical understanding, logical reasoning, experiment automation, or protocol adherence. Because each task requires different capabilities and exhibits distinct failure patterns, LLM performance varies substantially across tasks. Accordingly, this study defines suitability as the degree to which an LLM can support a given research task in a stable and efficient manner. This notion extends beyond average accuracy and incorporates factors such as consistency, error characteristics, and the practical usability of outputs. Conversely, unsuitability refers to recurring behaviors that introduce errors or inefficiencies and hinder the research workflow within a specific task context.

2.4 Agent-Based Analysis

An agent in this study denotes an automated analysis entity that evaluates multiple LLMs under controlled conditions and systematically extracts recurring response patterns and task-level rules. By executing predefined prompts and datasets, collecting evaluation outputs, and conducting comparative analyses, the agent reduces manual effort while ensuring consistency and reproducibility across experiments. This agent-based approach enables scalable and systematic analysis of LLM behavior across diverse research tasks.

3 Methodology

The overall benchmark evaluation pipeline of this study is illustrated in Figure 1. First, input datasets are prepared using MMLU, GSM8K, HumanEval, and IFEval, with sample sizes of n = 25, 20, 15, and 10, respectively. To ensure fair comparison, the same prompt template and decoding settings are fixed across all models. Inference is then performed using multiple LLMs, including DeepSeek, Qwen, Llama, EXAONE, and TinySwallow. Model outputs are evaluated with task-specific evaluators: Accuracy is used for MMLU, Exact Match or Accuracy for GSM8K, Pass@1 for HumanEval, and either Strict or Loose compliance rates for IFEval. Finally, the resulting metrics are used to compare model performance and to analyze task-specific failure patterns across models.

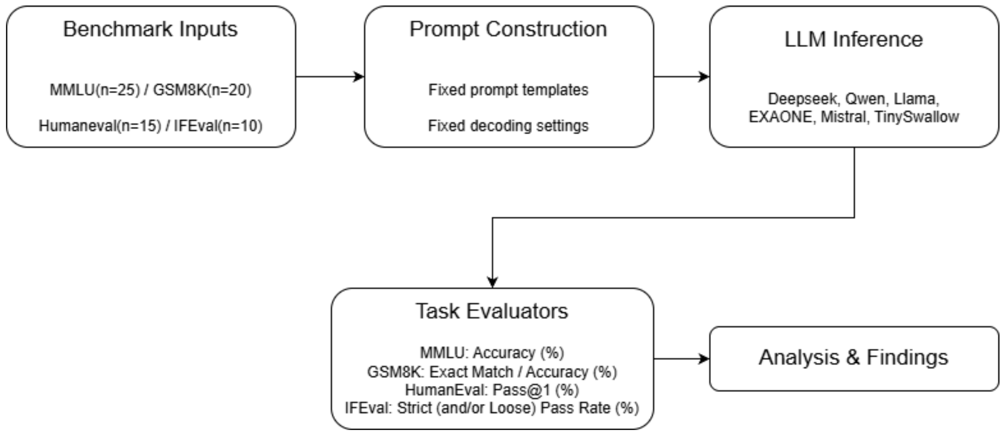

Figure 1: Overall evaluation pipeline.

3.1 Input Data

This study evaluates the suitability of large language models (LLMs) as research collaborators across different research task types using publicly available benchmarks and models from diverse families. All models are evaluated under identical input data, prompt templates, and decoding settings to ensure reproducibility and fair comparison. Six LLMs representing distinct development backgrounds are selected based on public accessibility, diversity of model families, and applicability to a broad range of research tasks, with the analysis focusing on task-level response characteristics rather than absolute performance ranking. Four benchmark categories are employed to reflect different research task requirements: multi-domain multiple-choice benchmarks for background knowledge and conceptual reasoning, mathematical reasoning benchmarks for step-by-step logical consistency, code generation benchmarks for experimentation and simulation tasks, and instruction-following benchmarks for protocol adherence. To control computational cost and enable repeated experiments, a subset of queries is sampled from each benchmark with balanced difficulty and topic coverage. All queries are embedded in a common prompt template without model-specific instructions, ensuring that observed differences primarily reflect inherent model behavior.

3.2 Task-Benchmark Mapping Framework

This study adopts the assumption that the capability dimensions measured by existing LLM benchmarks partially correspond to the core requirements of research tasks encountered in practical research workflows. Rather than treating benchmark scores as indicators of overall model superiority, we interpret them as reflecting task-relevant capability characteristics. Based on this assumption, we construct a task–benchmark mapping framework that links benchmark evaluation capabilities to representative types of research tasks, which serves as the analytical basis for subsequent suitability analysis.

In this study, a research task is defined as a goal-oriented activity performed at a particular stage of the research process, characterized by the primary cognitive and technical abilities it requires. Based on an analysis of common research workflows in which LLMs are frequently used, research tasks are categorized into four types: (1) theoretical and background knowledge–based tasks, which involve conceptual understanding and integration of prior work; (2) logical and mathematical reasoning–based tasks, which require step-by-step reasoning and logical consistency; (3) experimentation and simulation–based tasks, which center on specification-driven code generation and experimental automation; and (4) research protocol compliance tasks, which require strict adherence to output formats and constraints.

Existing benchmarks emphasize different capability dimensions corresponding to these task types. Multi-domain multiple-choice benchmarks primarily assess background knowledge and conceptual reasoning. Mathematical reasoning benchmarks evaluate logical coherence in step-by-step problem solving. Code generation benchmarks focus on the correctness of implementations based on explicit specifications, reflecting experiment-oriented research tasks. Instruction-following benchmarks emphasize compliance with output formats and constraints required by research protocols. Accordingly, each benchmark is associated with the research task type whose primary capability requirements it most strongly reflects.

The resulting task–benchmark mapping defines an approximate correspondence between benchmarks and research task categories, rather than a one-to-one or substitutive relationship. A single research task may require multiple capabilities, and a benchmark may partially capture several ability dimensions. In this study, the mapping is used as an analytical reference for interpreting benchmark results at the task level, enabling comparison of relative response tendencies and stability across model families. Based on this framework, the following section describes how task-level suitability and unsuitability criteria are derived from empirical experimental observations.

3.3 Deriving Task-Level Suitability and Unsuitability Rules

This study employs a structured analysis procedure to summarize and compare LLM response characteristics across different research task types under controlled execution conditions. In this context, "rules" refer to empirically derived criteria that organize observed performance metrics and response patterns from a task-oriented perspective, rather than outcomes produced through automatic inference or learning. All evaluated LLMs are executed using the same benchmark query sets and a shared prompt template. Each model is run independently under identical input conditions and configured to generate a single response per query. No model-specific instructions, output post-processing, or human intervention are applied. Generated responses are stored in a structured format together with metadata such as model identifiers, task categories, and execution conditions, ensuring that observed differences can be attributed to model characteristics or task suitability. Responses are grouped by research task type according to the predefined task–benchmark mapping. For each task category, quantitative performance indicators (e.g., accuracy or success rate) are summarized on a per-model basis, and qualitative response characteristics are recorded, including output format compliance, specification violations, and stability of reasoning. The analysis emphasizes recurring patterns observed across multiple instances within the same task category, rather than isolated successes or failures. Based on these task-level summaries, the study derives suitability and unsuitability criteria that describe relative tendencies of model families for specific research tasks. Suitability criteria capture cases in which models exhibit comparatively stable and effective response behavior, while unsuitability criteria highlight recurring failure patterns or inefficiencies that may disrupt research workflows. These criteria are derived through comparative analysis across models and task types and are intended to support task-aware interpretation of benchmark results, rather than to assert absolute performance rankings or universal judgments.

3.4 Form of Outputs

The proposed agent-based framework produces structured outputs that summarize LLM response characteristics at the research task level. For each task category, the framework derives task-level suitability and unsuitability rules based on recurring response patterns observed under controlled experimental conditions, capturing relative stability, effectiveness, and common failure modes of different model families. In addition, auxiliary task-level indicators are generated to provide quantitative context for these rules and support descriptive analysis in the experimental results section. All outputs are organized in a structured format indexed by research task category and model identifier, enabling consistent comparison across task types and experimental settings.

4 Experimental Design

This section describes the experimental design and evaluation methodology used to analyze how large language models exhibit different utilization characteristics across research task types. The purpose of the experiments is to examine, in a descriptive manner, how task-based criteria organized by an agent align with the actual task performance observed under controlled experimental settings. In addition, the analysis considers whether these criteria maintain consistent judgment tendencies under identical or similar experimental conditions. Accordingly, the evaluation is conducted from three perspectives: descriptive alignment, practical referential usefulness, and reproducibility and stability within a limited scope [8], [28]. All experiments are conducted under controlled conditions to ensure that observed differences can be attributed to model characteristics or task properties, rather than to variations in prompt design or input format [29].

4.1 Experimental Setup

The experiments are conducted on six large language models. Each model is evaluated using benchmark datasets corresponding to four research task types. Identical problem sets and the same prompt templates are applied to all models in order to minimize the influence of prompt phrasing or input format differences [30], [31]. The number of problems for each benchmark is fixed in advance, taking into account problem structure and evaluation cost, and this configuration is kept consistent across all experimental runs. Model outputs are stored in a structured format along with metadata such as task type, benchmark identifier, model identifier, and execution conditions [32]. The agent does not intervene in the output generation process of the models and operates solely as a post hoc analysis layer that aggregates, organizes, and compares the collected results. This design ensures a clear separation between the response generation process and the analysis process [33].

4.2 Alignment Analysis Between Task-Based Criteria and Performance Characteristics

The first evaluation focuses on whether the task-based criteria organized by the agent show alignment with actual research task performance outcomes. To this end, models are grouped into those that exhibit relatively suitable characteristics for each task type and those that do not. The analysis compares task-level performance distributions and response characteristics observed across these groups [34], [35]. Rather than focusing on correctness at the individual problem level or on single performance rankings, the evaluation emphasizes overall performance tendencies and response stability that are repeatedly observed at the task-type level. This approach allows the analysis to assess whether the organized criteria capture interpretable, task-oriented patterns, rather than outcomes that depend on specific problem selections.

4.3 Practical Referential Value of Unsuitability Criteria

The second evaluation examines whether the unsuitability criteria provide practically useful reference information for identifying failure patterns or inefficient response behaviors that recur during actual research tasks. For this purpose, model-task combinations that exhibit relatively unsuitable characteristics for certain task types are identified, and the failure modes and response properties observed in these combinations are summarized [36], [37]. Failure patterns are categorized into types that directly affect task execution, including output format violations, instruction noncompliance, incomplete code generation, and instability in reasoning processes [25]. By comparing the frequency with which these failure types are repeatedly observed, the analysis describes whether the unsuitability criteria can serve as a reference for researchers to anticipate potentially inefficient model-task combinations in advance.

4.4 Reproducibility and Stability of Judgment Tendencies

The final evaluation examines whether the organized criteria maintain overall judgment tendencies without excessive dependence on a specific experimental configuration. To this end, repeated runs are conducted under the same problem sets and experimental settings, while applying limited condition variations such as minor changes in prompt phrasing or differences in model versions [38]. The evaluation does not require complete identity of outcomes. Instead, it focuses on whether task-level model categorization tendencies and core response characteristics are generally preserved. This analysis does not aim to claim statistical robustness or broad generalization, but rather to descriptively summarize the stability of judgment tendencies observed within the limited experimental scope [39].

5 Experimental Results

5.1 Quantitative Results Analysis

The quantitative performance results across research task types are summarized in Table 1. The table reports task-level performance metrics for six large language models evaluated on four benchmarks corresponding to distinct research task categories. These results are analyzed descriptively to examine how model performance varies across task types, rather than to establish an overall performance ranking.

Substantial performance divergence is observed across benchmarks associated with different research tasks. On reasoning-oriented benchmarks, performance varies widely among models. For example, GSM8K accuracy ranges from 0.0% (Qwen) to 75.0% (Llama), with Exaone achieving 70.0% and Mistral achieving 35.0%. In contrast, TinySwallow, which attains the highest accuracy on MMLU (52.0%), exhibits comparatively lower performance on GSM8K (20.0%), indicating that background knowledge and multi-step reasoning capabilities are not consistently aligned across models.

On experimentation- and simulation-oriented tasks measured by HumanEval, a different performance pattern emerges. Mistral and TinySwallow achieve the highest Pass@1 scores (93.3%), followed by Qwen (86.7%) and Llama (73.3%), while DeepSeek and Exaone record 0.0% Pass@1. These results demonstrate that strong reasoning performance does not necessarily translate to effective code generation behavior, and vice versa.

For protocol-adherence tasks evaluated using IFEval, all models exhibit 0.0% pass rates under the fixed experimental setting. This uniform outcome indicates a shared limitation across evaluated model families in maintaining strict output constraints, suggesting that protocol compliance remains a challenging aspect of current LLM behavior irrespective of overall capability differences.

Across all evaluated models, no single model consistently outperforms others across all task categories. Instead, each model exhibits task-specific strengths and weaknesses, with performance differences often exceeding 40–70 percentage points between task types for the same model. These task-dependent performance gaps are larger and more systematic than variability observed at the individual problem level, indicating that task-level performance tendencies provide more informative signals than isolated instance-level outcomes.

Table 1: Task-level benchmark performance across models (higher is better)

Model	MMLU(Acc, %)	GSM8K(Acc, %)	IFEval (Pass/Score, %)	HumanEval (Pass@1, %)
DeepSeek	24.0	30.0	0.0	0.0
Qwen	20.0	0.0	0.0	86.7
Mistral	20.0	35.0	0.0	93.3
Llama	20.0	75.0	0.0	73.3
Exaone	20.0	70.0	0.0	0.0
TinySwallow	52.0	20.0	0.0	93.3

5.2 Failure Cases and Validation of Unsuitability Criteria

This analysis focuses on model task combinations in which relatively unsuitable characteristics were observed for specific research tasks, examining failure cases and inefficient output patterns that repeatedly appeared during the experiments. The analysis identified recurring failure types in these combinations, including violations of output format constraints, failure to follow instructions, incomplete code generation, and instability in reasoning processes.

Such failure patterns were not limited to reduced accuracy but were often associated with situations where researchers were required to perform additional corrections or repeated executions. In contrast, for model task combinations where unsuitable characteristics were not prominent, outputs tended to maintain greater structural stability and continuity in task execution. These comparative results suggest that the unsuitability criteria are not intended as absolute exclusion rules for specific models, but rather as conservative reference indicators to help identify combinations with a relatively high risk of failure during research workflows.

5.3 Reproducibility and Sensitivity Analysis

This section examines whether the organized task-based criteria maintain overall judgment tendencies under limited variations in experimental conditions, rather than being overly dependent on specific settings or incidental factors. To this end, repeated experiments were conducted using the same problem sets and base configurations, while introducing limited variations such as minor changes in prompt phrasing or differences in model versions. The results show that while some variations were observed in the detailed expressions or scope of the criteria, the overall model classification tendencies by research task type and the main failure signals were largely preserved. This indicates that the proposed criteria reflect relatively stable judgment tendencies within a constrained experimental setting, rather than being artifacts of specific experimental conditions. This analysis does not claim statistical robustness or generalizability, but instead aims to descriptively summarize the consistency of observed judgment patterns under controlled variations.

5.4 Human Evaluation Results

An optional human evaluation was conducted to assess how understandable and usable the task-based criteria are for actual researchers. Participants with research experience evaluated the criteria with respect to readability, ease of interpretation, and usefulness as a reference in research task selection. The evaluation results indicate that participants found the task-based criteria more helpful for context-aware decision making than single performance rankings. In particular, the unsuitability criteria were evaluated as useful reference information for identifying model task combinations with a higher likelihood of failure in advance. However, some participants noted as a limitation that the research task type must be clearly defined in order to apply the criteria effectively.

6 Discussion and Limitations

This study analyzed the behavior of large language models (LLMs) across different research task types using a task-oriented interpretation of public benchmark results. The quantitative results in Table 1 show that model performance varies substantially across task categories and that no single model exhibits uniformly strong behavior across all research-relevant tasks. For example, models that achieved relatively high accuracy on reasoning-oriented benchmarks such as GSM8K (e.g., Llama at 75.0% and Exaone at 70.0%) exhibited markedly lower performance on code generation tasks measured by HumanEval, where their Pass@1 scores dropped to 73.3% and 0.0%, respectively. Conversely, models with strong HumanEval performance, such as Mistral and TinySwallow (both at 93.3%), showed comparatively lower or inconsistent accuracy on reasoning-oriented benchmarks. These discrepancies indicate that aggregate rankings obscure task-dependent performance differences that are directly relevant to research workflows.

Beyond absolute performance values, the analysis revealed consistent differences in response stability and failure patterns at the task level. For instruction-following and protocol-adherence tasks (IFEval), all evaluated models recorded 0.0% strict or loose pass rates under the fixed experimental setting, indicating systematic difficulty in maintaining output constraints across models rather than isolated model-specific failures. Such uniform failure signals suggest that certain benchmark-task combinations expose structural limitations shared across current LLM families, reinforcing the need to interpret benchmark outcomes in relation to task requirements rather than as isolated capability scores.

A distinguishing analytical contribution of this study is the explicit examination of unsuitability patterns alongside relative strengths. While prior work often emphasizes identifying the best-performing model, the task-level analysis in this study documents model–task combinations that repeatedly produced incomplete code, format violations, or unstable reasoning traces across repeated runs. These unsuitability patterns were observed consistently within specific task categories, even when absolute accuracy differed across models [41], [42]. From an analytical perspective, such recurring failure tendencies provide complementary information to performance scores by highlighting conditions under which model outputs are more likely to disrupt research workflows.

The agent employed in this study functions solely as a post hoc analytical mechanism that aggregates and organizes observed outputs under identical experimental conditions. It does not participate in model response generation, scoring decisions, or task execution. All evaluations were conducted using fixed prompts, deterministic decoding (temperature = 0), and publicly defined metrics, ensuring that observed differences arise from model behavior rather than stochastic sampling effects or prompt variation [43]. As a result, the reported patterns reflect deterministic response tendencies under controlled conditions rather than probabilistic performance estimates.

Several limitations and threats to validity should be considered when interpreting the results.

First, the analysis is based on a limited number of benchmark samples (e.g., 25 for MMLU, 20 for GSM8K, and 15 for HumanEval), which constrains statistical generalization. The study does not claim statistical significance or population-level inference; instead, it focuses on identifying repeated task-level patterns that remain stable across controlled executions.

Second, the task–benchmark mapping relies on approximate correspondence between benchmark capability dimensions and real research tasks. While benchmarks such as GSM8K and HumanEval capture important aspects of reasoning and implementation, they cannot fully represent the complexity of end-to-end research activities [9], [44].

Third, the use of deterministic decoding eliminates variance due to sampling but also limits analysis of stochastic robustness. This choice was intentional, as the goal of the study is descriptive comparison of stable response tendencies rather than estimation of expected performance distributions.

Fourth, the evaluated model set represents a subset of publicly accessible LLM families and does not include proprietary or rapidly evolving models, which may exhibit different task-level characteristics.

Finally, the human evaluation component was limited in scale and scope and does not assess long-term productivity or collaborative efficiency in real research environments.

Despite these limitations, the observed task-dependent performance gaps, uniform failure signals on protocol-adherence tasks, and consistent unsuitability patterns across repeated runs suggest that the findings are not incidental artifacts of individual benchmarks or models. Rather, they provide empirical evidence that LLM behavior in research contexts is strongly conditioned on task characteristics. By explicitly documenting these task-level tendencies and limitations, this study complements performance-centric evaluation approaches and provides an analytical basis for further investigation into task-aware interpretation of LLM benchmark results.

7 Conclusion

This study examined the use of large language models (LLMs) in research from a task-oriented analytical perspective, focusing on how model response characteristics vary across different types of research tasks. Rather than interpreting LLM evaluation as a problem of single-score performance comparison, the analysis treated benchmark results as descriptive signals that reflect task-dependent behavior under controlled experimental conditions. By reinterpreting public benchmark outcomes through an explicit task–benchmark mapping and organizing performance distributions and response patterns observed in repeated experiments, the study aimed to characterize LLM usage patterns in a structured manner.

The empirical analysis shows that LLM behavior differs substantially across research task types and that overall benchmark scores or aggregate rankings alone do not sufficiently describe model behavior in research contexts. In particular, recurring failure patterns and stability characteristics were observed to depend on the interaction between task requirements and model response tendencies. These observations suggest that model suitability in research settings is context-dependent and cannot be inferred solely from absolute performance levels.

This work positions AI systems not as substitutes for researcher judgment, but as objects of analysis whose behavior can be examined under reproducible conditions. By organizing task-level response tendencies and limitations in a structured and interpretable form, the study provides descriptive reference information that may assist researchers in understanding potential model–task mismatches and in reducing uncertainty during model selection.

Future work may extend this analytical perspective by considering a broader range of research tasks, benchmark types, and model families, as well as by examining how task-oriented response characteristics evolve under different prompting strategies or experimental settings. Such extensions could further clarify the scope and limitations of task-oriented interpretations of LLM behavior in research workflows.

References

[1] Boiko, D.A., MacKnight, R., Kline, B. & Gomes, G. (2023) Autonomous chemical research with large language models. Nature 624(7992):570–578.

[2] Liu, Z., Chai, Y. & Li, J. (2024) Towards fully autonomous research powered by LLMs: Case study on simulations. arXiv preprint arXiv:2408.15512.

[3] Ma, P., Wang, T.-H., Guo, M., et al. (2024) LLM and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. arXiv preprint arXiv:2405.09783.

[4] Benegas, G., Ye, C., Albors, C., Li, J.C. & Song, Y.S. (2025) Genomic language models: Opportunities and challenges. Trends in Genetics.

[5] Ting, Y.-S., Nguyen, T.D., Ghosal, T., et al. (2024) AstroMLab 1: Who wins astronomy Jeopardy!? Astronomy and Computing:100893.

[6] Beeching, E., Fourrier, C., Habib, N., et al. (2023) Open LLM leaderboard. Online resource.

[7] Srivastava, A., Rastogi, A., Rao, A., et al. (2022) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

[8] Bommasani, R., Liang, P. & Lee, T. (2023) Holistic evaluation of language models. Annals of the New York Academy of Sciences 1525(1):140–146.

[9] Lin, B.Y., Deng, Y., Chandu, K., et al. (2024) WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. arXiv preprint arXiv:2406.04770.

[10] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. & Steinhardt, J. (2021) Measuring massive multitask language understanding. In Proc. ICLR 2021.

[11] Hendrycks, D., Burns, C., Kadavath, S., et al. (2021) Measuring mathematical problem solving with the MATH dataset. In Proc. NeurIPS 2021.

[12] Chen, Y., Arkin, J., Dawson, C., Zhang, Y., Roy, N. & Fan, C. (2024) AutoTAMP: Autoregressive task and motion planning with LLMs as translators and checkers. In Proc. IEEE International Conference on Robotics and Automation (ICRA), pp. 6695–6702.

[13] Madireddy, S., Xu, C., Cappello, F., et al. (2024) Comprehensive multi-stage evaluation of language models for scientific skill and safety red-teaming. In TPC Workshop at SC24 (Workshop).

[14] Underwood, R., Madhyastha, M., Burns, R. & Nicolae, B. (2024) Evostore: Towards scalable storage of evolving learning models. In Proc. HPDC 2024.

[15] McNaughton, A.D., Ramalaxmi, G., Kruel, A., et al. (2024) CACTUS: Chemistry agent connecting tool-usage to science. arXiv preprint arXiv:2405.00972.

[16] Schick, T., Dwivedi-Yu, J., Dessì, R., et al. (2023) Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems 36:68539–68551.

[17] Hong, Z.-W., Shenfeld, I., Wang, T.-H., et al. (2024) Curiosity-driven redteaming for large language models. In Proc. ICLR 2024.

[18] Achiam, J., Adler, S., Agarwal, S., et al. (2023) GPT-4 technical report. arXiv preprint arXiv:2303.08774.

[19] Anthropic (2025) Introducing Claude 4. Web page (accessed 2025-09-22).

[20] Anthropic Team (2024) The Claude 3 model family: Opus, Sonnet, Haiku. Papers With Code.

[21] Grattafiori, A., Dubey, A., Jauhri, A., et al. (2024) The Llama 3 herd of models. arXiv preprint arXiv:2407.21783.

[22] Qwen Team (2025) Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923.

[23] Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021) Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

[24] Chen, W., Zha, H., Chen, Z., Xiong, W., Wang, H. & Wang, W.Y. (2020) HybridQA: A dataset of multi-hop question answering over tabular and textual data. Findings of EMNLP 2020.

[25] Hua, T., Hua, H., Xiang, V., et al. (2025) ResearchCodeBench: Benchmarking LLMs on implementing novel machine learning research code. arXiv preprint arXiv:2506.02314.

[26] Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C. & Tafjord, O. (2018) Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457.

[27] Duan, J., Cheng, H., Wang, S., et al. (2024) Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In Proc. ACL 2024, pp. 5050–5063.

[28] Wang, Y., Ma, X., Zhang, G., et al. (2024) MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. arXiv preprint arXiv:2406.01574.

[29] Gao, L., Tow, J., Abbasi, B., et al. (2024) A framework for few-shot language model evaluation. Zenodo (record 12608602).

[30] Lin, Z., Trivedi, S. & Sun, J. (2023) Generating with confidence: Uncertainty quantification for black-box large language models. Transactions on Machine Learning Research.

[31] Wei, J., Wang, X., Schuurmans, D., et al. (2022) Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35:24824–24837.

[32] Kwon, W., Li, Z., Zhuang, S., et al. (2023) Efficient memory management for large language model serving with PagedAttention. In Proc. SOSP 2023, pp. 611–626.

[33] Albert, D. & Billinger, S. (2024) Reproducing and extending experiments in behavioral strategy with large language models. arXiv preprint arXiv:2410.06932.

[34] Li, L., Dong, B., Wang, R., et al. (2024) SALAD-Bench: A hierarchical and comprehensive safety benchmark for large language models. arXiv preprint arXiv:2402.05044.

[35] Zhang, Z., Jiang, Z., Xu, L., Hao, H. & Wang, R. (2024) Multiple-choice questions are efficient and robust LLM evaluators. arXiv preprint arXiv:2405.11966.

[36] Zhu, Z., Yang, Y. & Sun, Z. (2024) HaluEval-Wild: Evaluating hallucinations of language models in the wild. arXiv preprint arXiv:2403.04307.

[37] Du, X., Xiao, C. & Li, Y. (2024) HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In Proc. NeurIPS 2024 (OpenReview).

[38] Zhang, H., Da, J., Lee, D., et al. (2024) A careful examination of large language model performance on grade school arithmetic. arXiv preprint arXiv:2405.00332.

[39] Xiong, M., Hu, Z., Lu, X., et al. (2024) Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In Proc. ICLR 2024.

[40] Wang, B., Chen, W., Pei, H., et al. (2024) DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. Advances in Neural Information Processing Systems 36.

[41] Buszydlik, A., Dobiczek, K., Okon, M.T., et al. (2023) Red teaming for large language models at scale: Tackling hallucinations on mathematics tasks. arXiv preprint arXiv:2401.00290.

[42] Shen, X., Chen, Z., Backes, M., Shen, Y. & Zhang, Y. (2024) "Do Anything Now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. arXiv preprint arXiv:2308.03825.

[43] Yang, Z., Qi, P., Zhang, S., et al. (2018) HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proc. EMNLP 2018.

[44] Guo, T., Nan, B., Liang, Z., et al. (2023) What can large language models do in chemistry? A comprehensive benchmark on eight tasks. Advances in Neural Information Processing Systems 36:59662–5968

AI Co-Scientist Challenge Korea Paper Checklist

1. **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The claims in the abstract and introduction accurately reflect the scope of this work, which focuses on task-oriented reinterpretation of benchmark results and descriptive analysis of LLM suitability rather than predictive modeling or optimization.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors? Answer: [Yes]

Justification: The paper explicitly discusses limitations related to benchmark coverage, fixed prompt settings, and the descriptive nature of the derived rules in the Discussion and Limitations sections.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an impor- tant role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not present formal theoretical results, theorems, or proofs, as it focuses on empirical analysis and descriptive reasoning.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main ex- perimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All experimental settings, benchmarks, prompts, and evaluation procedures are fully described in the paper, enabling reproduction of the reported results using the same models and datasets.

Guidelines:

- The answer NA means that the paper does not include experiments. If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instruc- tions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: While the experiments rely entirely on publicly available benchmarks and models, the evaluation code is not released as part of this submission.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable). Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper- parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all relevant experimental details, including benchmark selection, task-to-benchmark mapping, prompt design, model configurations, and evaluation criteria.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Statistical significance tests and error bars are not reported because the experiments use deterministic inference settings (temperature = 0) and focus on descriptive task-level performance trends rather than stochastic variability.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the type of models evaluated and the general computational environment used for inference, which is sufficient for reproducing the experimental runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://nips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research adheres to the NeurIPS Code of Ethics by using publicly available datasets and models, preserving anonymity, and avoiding any form of human subject experimentation.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consid- eration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both positive impacts, such as improved decision-making in LLM usage, and potential negative impacts, including over-reliance on benchmark-driven model selection, along with appropriate cautions.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work does not release new models, datasets, or assets that pose risks of misuse, and therefore no additional safeguards are required.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer:[Yes]

Justification: All benchmarks and models used in the study are publicly available, properly cited, and used in accordance with their respective licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release any new datasets, models, or software assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The study does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribu- tion of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects are involved in this research, so IRB approval is not required.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.