

Assessing and Verifying Task Utility in LLM-Powered Applications

Negar Arabzadeh^{1*} Siqing Huo¹ Nikhil Mehta² Qingyun Wu³ Chi Wang⁴
Ahmed Awadallah⁴ Charles L. A. Clarke¹ Julia Kiseleva⁴

¹University of Waterloo, ²Purdue University, ³Pennsylvania State University,
⁴Microsoft Research

Abstract

The rapid development of Large Language Models (LLMs) has led to a surge in applications that facilitate collaboration among multiple agents, assisting humans in their daily tasks. However, a significant gap remains in assessing to what extent LLM-powered applications genuinely enhance user experience and task execution efficiency. This highlights the need to verify utility of LLM-powered applications, particularly by ensuring alignment between the application’s functionality and end-user needs. We introduce AgentEval, a novel framework designed to simplify the utility verification process by automatically proposing a set of criteria tailored to the unique purpose of any given application. This allows for a comprehensive assessment, *quantifying the utility of an application against the suggested criteria*. We present a comprehensive analysis of the effectiveness and robustness of AgentEval for two open source datasets including Math Problem solving and ALF-World House-hold related tasks. For reproducibility purposes, we make the data, code and all the logs publicly available at <https://github.com/Narabzad/AgentEval/>

1 Introduction

One of the long-lasting goals for intelligent agents (Winograd, 1972) is for them to seamlessly interact with humans in natural language and help their end-users with their tasks, such as completing household tasks, math tutoring, and so on. The rapid development of open-source libraries (Wu et al., 2023; Li et al., 2023a) helps that goal by simplifying the development of LLM-powered agentic applications for various user-centered tasks (Liang et al., 2023b; Hong et al., 2023; Talebirad and Nadiri, 2023; Arabzadeh et al., 2022; Mohanty et al., 2024). To ensure that the application’s behavior meets the requirements of the application

* Work done during an internship at Microsoft Research

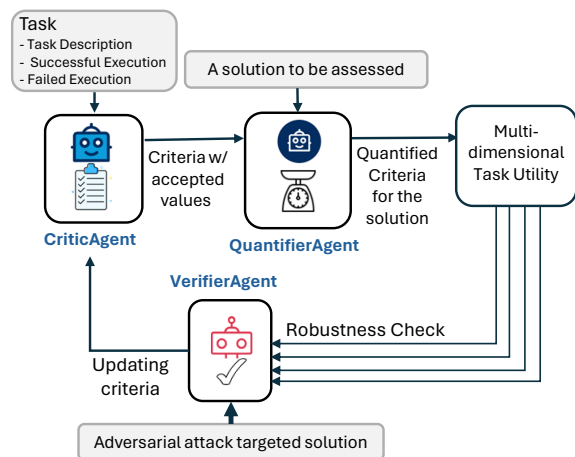


Figure 1: An overview of the *AgentEval* framework: *CriticAgent* creates a set of criteria and suggested values; *QuantifierAgent* quantifies the criteria for a considered application; and *VerifierAgent* verifies the criteria based on its robustness. The output of the *QuantifierAgent* is a multi-dimensional assessment of the utility of the application based on a suggested list of criteria and their evaluations.

developers, it is also crucial to assess its potential utility to end users (Dibia et al., 2023; Nguyen et al., 2016), as this can significantly impact its improvement journey. Taking into account a range of applications, it is unrealistic to assume benchmarking for every domain, including but not limited to code generation (Liu et al., 2024), health care (Andrew, 2024), and many others whose development we witness every day (Wu et al., 2023). Moreover, directly evaluating agentic applications poses challenges, as current approaches predominantly rely on end-to-end success metrics i.e., whether the application accomplishes tasks (Shridhar et al., 2020b, 2019; Myers et al., 2023). However, understanding a user’s interactions with an application involves much more than success alone (Kiseleva et al., 2022a,b; Zhang et al., 2023). Consider math problem solving, although it is important that the application solves the problem correctly, its ability to present and explain solutions based on various

criteria, such as completeness, conciseness, and clarity, is crucial. Furthermore, success is not always clearly defined for a task. Recognizing such criteria and being able to quantify them is essential to assess whether developer requirements are being satisfied and if the application brings utility to the end-users. Given the objective of assessing arbitrary applications, relying solely on end-to-end success metrics is untenable, due to the expansive range of tasks requiring automation. The question is *how to design a flexible methodology to assess the task utility for diverse set of applications?*

To bridge this gap, we introduce *AgentEval*, a framework to gauge the utility of LLM-powered applications. Its goal is to assess the utility by providing application developers with insights into how the current flow can be characterized. *AgentEval* builds on recent work showing that LLMs can be a scalable and cost-effective alternative to human evaluation for open-ended tasks (Li et al., 2023b). *AgentEval* as illustrated in Fig. 1, consists of the three following agents, formally defined in Sec. 3: (1) *CriticAgent* suggests the list of criteria based on the task description and a pair of solutions, where one is preferred over the other one (e.g., successful and failed examples). For instance, for math problems, the criteria could be Efficiency and Clarity of the proposed solution; (2) *QuantifierAgent* quantifies how the solution performs for each criterion and returns the utility function, e.g. for math problems, if the 'Clarity is 'not clear', 'moderately clear', or 'very clear'; (3) *VerifierAgent* verifies the quality of the assessment of the suggested criteria to make sure the criteria are essential, robust, informative and have high discriminative power.

In summary, our main contributions are **C1** Introducing *AgentEval*, a novel framework that leverages LLM-powered agents as a scalable and cost-effective alternative to human evaluations, to produce task utility through the collaboration of *CriticAgent*, *QuantifierAgent* and *VerifierAgent*; and **C2** An in-depth analysis of *AgentEval* robustness for two applications across different solutions, that can be replicated on an unseen domain.

2 Related Work

2.1 Evaluation of LLMs

Prior work (Guo et al., 2023; Ziyu et al., 2023; Chang et al., 2023; Liang et al., 2023a; Arabzadeh et al., 2024a) has extensively studied the evaluation

of LLMs on various fronts: how ethically sound they are (Stahl and Eke, 2024), how they align to human preferences (Hendrycks et al., 2021a; Köpf et al., 2024), their robustness (Wang et al., 2023b; Seifkar et al., 2023), and the knowledge, and reasoning capabilities they possess (Bian et al., 2023). Recent work evaluates LLMs on more specialized tasks, such as medical domain (Jin et al., 2019), multi-modal tasks (Mialon et al., 2023; Bang et al., 2023), or as agents in interactive environments (Liu et al., 2023).

2.2 User satisfaction prediction

Studies suggest that users interacting with various systems operate with specific utility functions in mind (Li et al., 2020; Azzopardi et al., 2018; Ahmadvand et al., 2022). Traditionally, metrics defining user satisfaction were designed using large-scale collected behavioral signals (Kiseleva et al., 2014), and were tailored to specific applications, such as intelligent assistants (Kiseleva et al., 2016a,b), web search engines (Williams et al., 2016a,b; Williams and Zitouni, 2017; Arabzadeh et al., 2023), dialogue systems (See et al., 2019), multi-turn conversations (Li et al., 2021; Mohanty et al., 2023) and general-purpose personal assistants (Kiseleva and de Rijke, 2017). It was demonstrated that assessing users' satisfaction requires goes beyond a single metric (Arabzadeh et al., 2024b). As such, here, we propose a flexible framework to assess user and developer requirements, which can eventually be used to improve the application flow.

2.3 Using LLMs as evaluators

More recently, there has been a growing trend in utilizing LLMs as evaluators (Chiang and Lee, 2023; Fu et al., 2023; Alaofi et al., 2024; Arabzadeh and Clarke, 2024; Huo et al., 2023,?), such as for qualitative research (Bano et al., 2023), or summarization. Specifically, Jain et al. (2023) studied the efficacy of few-shot prompted LLM evaluators in evaluating summaries that were written by other LLMs. Similarly, Wang et al. (2023a) explore if ChatGPT itself can be used as an evaluator, by prompting it to score texts. Other works (Tjuaatja et al., 2023; Liu and Sun, 2023; Chiang and Lee, 2023; Meng et al., 2024) look at how LLMs can be used as proxies for human behavior, or work with humans, such as CoEval (Li et al., 2023b), which showed how LLMs can make hu-

man evaluation easier. Pan et al. (2024) also show how LLM evaluators can help build models that increase performance on downstream task. Building on the above, a different line of works identify weaknesses in single LLMs as direct evaluators (Huang et al., 2023), and propose to improve them, such as a multi-step calibration framework (Wang et al., 2023c). Given these drawbacks, recent work has looked at how multiple LLM agents can be used as evaluators. Chan et al. (2023), proposed ChatEval, a multi-agent team that discusses and evaluates responses from agents on generation tasks (debate-style), leading to text that aligns with better human preferences. Similarly, Chern et al. (2024) proposed a multiple agent-debate-assisted meta-evaluation framework.

Building on these works, we propose an automatic multi-agent assessment of utility for arbitrary LLM-powered applications, to provide deep insights for developers. Our framework can uncover current flaws in these applications, and may lead to improvements in them, particularly if the application flow changes after it is applied, and then it is re-used.

3 Task Utility

Fig. 2 outlines a taxonomy of target tasks for LLM-powered applications, in terms of success metrics. At a high level, these tasks can be categorized into: 1) *Success is not clearly defined* — Users use the system in an assistive manner, seeking suggestions from it, rather than expecting it to solve the task. For example, a user can request the system to generate an email. The user usually uses the system’s response as a template, which can later be edited. Directly evaluating assistive tasks like these is hard, particularly for online evaluation, or when dealing with less well-defined tasks. One potential approach is to directly ask users how useful the help was, but this is not well-calibrated (Borisov et al., 2018), hard to quantify (Sepliarskaia et al., 2018), and expensive.

2) *Success is clearly defined* — It is clear whether the system solved the task or not, for example, assisting with household tasks, where success is clear and measurable. This category can be further divided into two subcategories:

- *an optimal solution exists* — only one successful outcome is possible. For example, when asking an assistant to turn on a light, success is clearly defined, as there is only one way to do it.

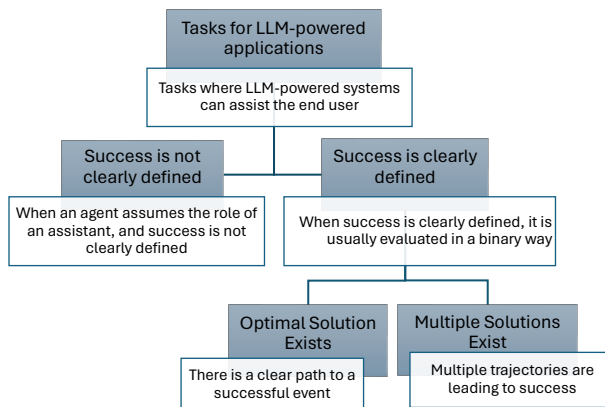


Figure 2: The taxonomy of tasks assessment.

- *multiple solutions exist* — Increasingly, we observe situations where multiple trajectories of agent behavior can lead to success. For example, when asking an agent to suggest a food recipe, success could be multiple cuisines tasting good, but perhaps the recipe should not be expensive.

AgentEval is currently focused on tasks where *success is clearly defined and multiple successful solutions* may exist.

Previous research on assistive agents suggests human pairwise preferences as one of the most optimal assessments, i.e. when the annotator is presented with two agents side by side and asked for their preferences (Kiseleva et al., 2022b). In this setup of side-by-side pairwise comparison, humans tend to suggest a list criteria, explaining why they prefer one agent over the other. For instance, ‘the first agent was faster’ or ‘the second agent converses more naturally’. This comparative setup can guide humans to come up with a list of criteria that helps to infer the utility of the task. With this in mind, we designed *AgentEval* (Fig. 1), by employing LLMs to help us understand, verify, and assess task utility, namely:

- *CriticAgent*: The goal of this agent is to suggest a set of criteria that can be used to assess task utility. The *CriticAgent* is given a task description, as well as optionally several pairs of solutions, where preferably some are preferred over the other ones, for instance, successful and failed examples. *CriticAgent* would return a set of criteria $C = \{c_1, \dots, c_n\}$, where each criterion c_i is accompanied by a set of accepted values ω as $c_i : \{\omega_j\}_{j=1}^m$. For example, for solving math problems, the *CriticAgent* generated accepted values and criteria such as clarity, efficiency, and more - see Tab. 1.
- *QuantifierAgent*: The goal of *QuantifierAgent*

is to quantify each of the suggested criterion, to access the task utility of the system U_t , for the end user. We define the Utility for task t as: $U_t(s) = \{Q_i(s|c_i)\}_{i=1}^n$. where s represents the task sample and $Q(s|c_i)$ is the quantifier output for sample s based on the criterion c_i . For example, for math problem solving, given the generated criteria shown in Tab. 1, the solution’s Accuracy could be quantified as “Incorrect”, “partially correct” or “correct”. Eligible quantified values for quantification process are shown in “Accepted values” column in Tab. 1

- *VerifierAgent*: There might be cases where not all the criteria suggested by *CriticAgent* help assess utility. Some criteria might be redundant, while others may not aid in distinguishing performance. *VerifierAgent* validates the quality of the criteria in terms of robustness and their distinguishability of noisy samples. Essentially, it checks (1) if the criteria can be quantified robustly over repeated samples, and (2) if *QuantifierAgent* can identify the adversarial attacked targeted samples from the original ones. If the sanity checks do not pass, *VerifierAgent* will update the list of criteria, to end up with a set of robust, stable, informative and distinguishable criteria for assessment.

Finally, we note that *AgentEval* allows for incorporating a human in the loop in the role of a domain expert. For instance, *CriticAgent* could be replaced by a human expert who either comes up with the relevant criteria or helps *VerifierAgent* verify the useful criteria and filter out the unessential ones.

4 Datasets and Solutions

This section provides an overview of the datasets utilized in our study i.e., Math problem solving and ALFWorld household task. The math dataset is chosen for its widespread usage and complex problem-solving scenarios that are fundamental in evaluating the effectiveness. ALFWorld dataset offers a scenario involving multi-turn interactions within a moderately approximated multi-modal environment. Each dataset plays a critical role in evaluating different aspects of *AgentEval*’s capabilities, from handling complex theoretical problems to navigating real-world scenarios. In both tasks, although success is clearly defined, multiple solutions exist for accomplishing the objectives. An example of Math problem solving and ALFWorld task is shown in Appendix A.1. Due to space, we report all experiments about Math problem solving

in the main paper and we keep all the experiments related to ALFWorld dataset in the Appendix A.3.

4.1 MATH Problem Solving

Dataset: The MATH dataset is a substantial collection of 12,500 challenging mathematics problems from high school competitions (Hendrycks et al., 2021b). Each problem comes with a step-by-step solution and is tagged by difficulty levels. Similar to the math problem experimental setup in Wu et al. (2023), we carry out evaluations on 120 problems from level-5 by three different solutions. Due to limited space, for more details about this dataset, we refer readers to Appendix A.2

Solutions: In establishing solutions for this task to assess, we draw inspiration from the experiments showcased in (Wu et al., 2023). We evaluate the proposed methodology by AutoGen (Wu et al., 2023), as well as Langchain ReAct (Yao et al., 2022) and a Vanilla solver that employs GPT-4 to tackle the task. These solutions have previously demonstrated promising and competitive performance (Wu et al., 2023). In Sec. 5.2, we explore how the measured performance with *AgentEval* correlates with the ground truths.

4.2 ALFWorld Household Task

Dataset: ALFWorld presents a set of language-based interactive decision-making tasks within simulated household environments (Shridhar et al., 2020b). ALFWorld is the first interactive parallel environment that aligns text descriptions and commands with physically embodied robotic simulation. Finally, the dataset’s inclusion of household chores to more intricate problem-solving scenarios, provides a comprehensive testbed for evaluating the adaptability of multi-agent systems. For more information about the dataset and examples of the test cases, we refer the readers to Appendix A.3.1.

Solutions: As for the solutions to assess for ALFWorld Household tasks, similar to (Wu et al., 2023), we consider ReAct (Yao et al., 2022) as well as AutoGen with two agents and AutoGen with three agents (Wu et al., 2023). In Appendix A.3.2, we discuss in more details the solutions under assessment. We assess and compare the performance of these three solutions using *AgentEval*.

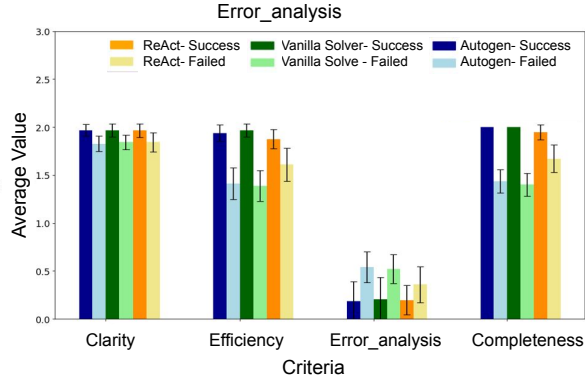


Figure 3: *AgentEval* assessment of three solutions on math problems categorized by success and failed cases.

5 Experiments

5.1 Implementation Details

For all experiments, we use GPT-4 version 0613, accessed through Azure OpenAI services, as the LLM model and the temperature of 0. *AgentEval* utilizes AutoGen (Wu et al., 2023) for implementation, since it provides a versatile environment where agents can be finely tuned and customized based on specific application needs. This is crucial for maintaining the flexibility to handle a wide range of applications. We tried to avoid much prompt engineering and tried to keep each agent’s instructions as if we are instructing human annotators. Moreover, another advantages of using AutoGen for implementation of *AgentEval* is that it has the flexibility to involve human in the loop. Each agent could be replaced by a human annotator. We further provide all the prompts used in our experiments in our Git repository.

5.2 AgentEval for Math Problems

When executing the *CriticAgent* for Math problem solving, we first obtain a set of criteria as presented in Tab. 1. Then, the *QuantifierAgent* is tasked with quantifying each criterion, based on the accepted values. We present the outcome of *QuantifierAgent* measuring performance of three solutions on this task in Fig. 3. Notably, we see that *Agenteval* does not quantify the three solutions as if they perform equally well across the different criteria. For instance, while all three solutions leverage GPT-4 as the underlying language model, AutoGen outperforms ReAct and Vanilla GPT-4 in terms of accuracy. This observation, while confirmed by previous studies (Wu et al., 2023), extends to solution completeness and efficiency as well.

As depicted in Fig. 3, the error analysis range of quantified values differs from other metrics. We

Table 1: Verification Criteria for MathProblems

Criteria	Description	Accepted Values
Clarity	The ease of understanding the steps, explanations, and language used in the solution.	– Not Clear (0) – Moderately Clear (1) – Very Clear (2)
Efficiency	The use of optimal methods or approaches to solve the math problem.	– Inefficient (0) – Moderately Efficient (1) – Efficient (2)
Error Analysis	The identification and description of possible errors or misconceptions in the math problem-solving process.	– Not Addressed (0) – Partially Addressed (1) – Well Addressed (2)
Completeness	Quality of code in terms of efficiency and elegance	– Incomplete (0) – Mostly Complete (1) – Complete (2)

scrutinize the results by categorizing them into successful and failed cases. AutoGen, Vanilla Solver and ReAct solutions are each presented in orange, blue and green respectively, where the darker bars represent the performance on successful cases and lighter bars represent the failed cases. The difference between the dark and light bar of each color, verify *AgentEval*’s performance, as we expect that each positive criteria should be quantified higher for successful cases compared to their failed cases. We observe that in most cases, the successful and failed cases are distinguished, even with 95% interval confidence on all the success and failed cases.

When examining the differences between successful and failed cases among the three solutions, we note that not all successful cases are assessed identically, nor are all failed cases quantified with the same performance. This can be interpreted to mean that even though two solutions might both be successful, one might perform better or worse in certain criteria, such as clarity or efficiency. This observation provides us with valuable additional insights, especially for the developers of the proposed solutions, and goes beyond reporting the effectiveness of a application by one scalar value e.g., success rate.

6 Robustness Analysis and Verification

In this section, we first analyze the robustness of *AgentEval*, then further investigate how *VerifierAgent* can increase the stability of our assessment.

6.1 Diversity of Criteria

Here, our main goal is to study the diversity of the suggested criteria. We investigate the extent inputs to *AgentEval* (Fig. 1 such as ‘Task Description’ and ‘Successful/Failed Executions’) contribute to *CriticAgent* for creating a more diverse set of criteria. To do so, we use two distinct methods, with *CriticAgent* generating (1) “task-based” criteria solely from the task description, and (2) “solution-based”

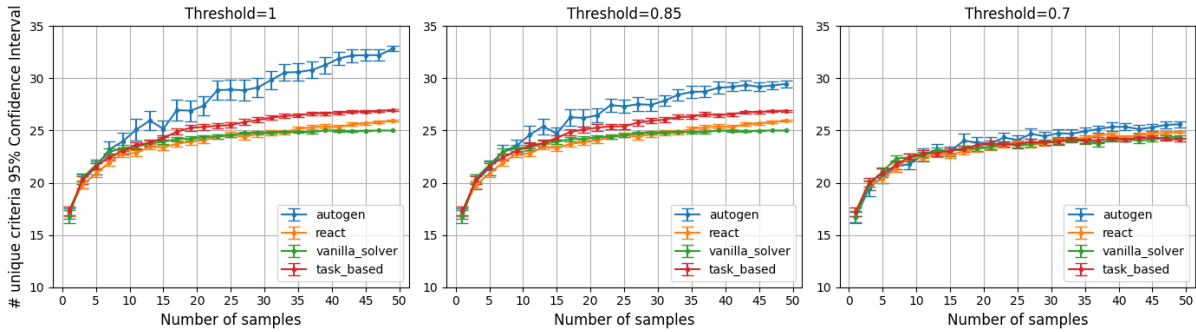


Figure 4: Task-based vs solution-based criteria for Math problems. Error bar show the 95% confidence interval.

criteria, derived from both the task and execution examples. For example, a solution to a mathematical problem, might satisfy criteria such as ‘Accuracy’ and ‘Clarity’, independent of the solution. However, when additional tools such as coding are used to solve the problems, additional criteria like ‘Code Efficiency’ may be introduced to the set of criteria. This makes sense, since the application leveraged coding to solve math problems.

Fig. 4 displays the number of unique criteria extracted for mathematical problem solving in task-based mode, and three different solution-based approaches. To keep the balance between computational costs and analyzing the robustness, we conducted 50 runs of the *CriticAgent* with different seeds. Subsequently, for $N = 50$ iterations, we randomly select $M \leq 50$ samples, as shown on the x-axis of Fig. 4, and present the average number of unique extracted criteria, along with its 95% confidence interval after repeating this process 50 times. We note that because the total pool of criteria includes 50 iterations in total, the confidence intervals become smaller when M get closer to the maximum number of samples i.e., 50

To gain deeper insights into diversity of criteria, we took a closer look at them to study if they are truly unique or to what extent they have similarities. This is important to determine if *CriticAgent*, when continually generating criteria, will always produce new criteria, or if it will eventually converge to a set. We noted that some criteria are similar but worded differently. For example, ‘Problem Complexity’ vs. ‘Problem Difficulty’ or ‘Time Taken’ vs. ‘Time to Completion’. Tab. 3 in the Appendix lists such instances. To consolidate the similar criteria and reduce noise in the number of unique criteria and redundancy, inspired from previous work (Liu et al., 2022; Vahtola et al., 2022; Reimers and Gurevych, 2019), we employ a pre-trained

language model fine-tuned for paraphrasing¹, to measure the semantic similarity of criteria descriptions. Using a threshold τ , we classify pairs with cosine similarity greater than τ as semi-identical ones and select one of them as the representative of the pair. Fig. 4 illustrates the impact of different τ values (0.7, 0.85, 1) on the diversity of criteria. A threshold of 1 means no filtering occurs. This analysis shows that the solution-based approach has potential to produce more diverse criteria than the task-based approach, although this varies by the creativity of the model. For example, while the AutoGen solution demonstrates the highest diversity, task-based methods yield more unique criteria than ReAct and Vanilla Solver. Another interesting observation is that repeating the *CriticAgent* will eventually lead to a convergence in the number of criteria. This suggests that the *CriticAgent*’s ability to create new criteria will diminish, converging to an almost finite list of criteria, which will reduce the cost as well.

6.2 Verification

As outlined in Sec. 3 and illustrated in Fig. 1, the *VerifierAgent*’s primary role is to ensure the selected criteria are effective toward evaluating the utility for the end-user, while maintaining robustness and high discriminative power. To achieve this, the *VerifierAgent* undertakes two main actions: (1) *Criteria Stability*: The criteria should be essential and robust, meaning they should not be redundant and we should be able to quantify them stably if we repeatedly quantify it for an individual solution, showing no divergence. As such, *VerifierAgent* enhances the criteria by iterating over the generation and quantification phases. It then consolidates these criteria by identifying and eliminating redundancies, followed by evaluating the

¹<https://bit.ly/3UgsY0p>

dispersion of the distribution of the quantified criteria. This step modifies the criteria, ensuring that only the most robust criteria are retained.

(2) *Discriminative Power*: A reliable evaluation should detect and withstand noise. To test that, we propose to use adversarial examples and then assess the system’s ability to differentiate between these compromised examples and standard cases. Should the system fail to distinguish effectively, it indicates that the criteria are insufficient for reliable assessment under varied conditions.

We note that both steps involve a tunable threshold that can be adapted based on application needs, ensuring flexible criteria validation. The proposed methodology for *VerifierAgent* is summarized in Algorithm 1 in the Appendix.

6.2.1 Criteria Stability

Our goal here is to explore the stability of criteria and robustness of the quantifier for having a more essential, robust and stable set of criteria. We specifically evaluate the *QuantifierAgent*’s robustness using criteria for mathematical problems (Table 1), conducting 50 repeats of runs with different seeds on 120 problems (Section 4.1). Ideal expected outcomes include consistent performance across all criteria on all the repeats. Fig. 5 illustrates the distribution of quantifier values for both failed (dark blue) and successful cases (light blue) across all criteria through box plots. The more robust a criterion, the narrower the range of quantified performance (narrower box plots). Also, the less overlap between the successful and failed boxes, the higher the distinguishability of the criteria. We observe that all four criteria, except ‘error analysis’ allow for easy differentiation between successful and failed cases. Additionally, some criteria prove to be more robust compared to others. We believe that such an analysis of the quantifier agent’s performance will yield valuable insights for enhancing reliability, trustworthiness, and explainability in performance evaluation. A detailed examination of the stability of each criterion, especially how they differentiate between successful and failed cases, is provided in Appendix A.4.2.

Further, to refine and expand the criteria set without redundancy, we operate the *CriticAgent* multiple times i.e., we execute *CriticAgent* 50 times with varied seeds. The criteria are then summarized into one list of useful criteria using the LLM. Additionally, as explained in Section 6.1, we remove similar and redundant criteria using pre-trained lan-

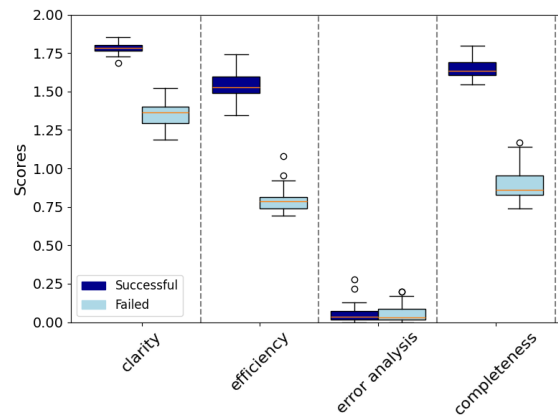


Figure 5: Distribution of *QuantifierAgent* output on AutoGen results on successful (dark blue) and failed (light blue) cases on different criteria.

guage models, thus obtaining a comprehensive list of criteria. The refined criteria after 50 repeats are detailed in Tab. 4 in the Appendix.

Now, we aim to determine the stability of these criteria through repeated quantifications. Our goal is to identify criteria that maintain consistent results without significant divergence, even when quantified multiple times. Using this consolidated list, we measure the dispersion of quantified results using the coefficient of variation, a standardized metric that facilitates comparison across various test cases when *QuantifierAgent* quantifies them. Given the consolidated list of criteria, we use the *QuantifierAgent* to quantify various test cases and report the coefficient of variation as a measure of the dispersion of the *QuantifierAgent*’s outputs with respect to each criterion across different seeds and report the mean coefficient of variation across all samples. we run *QuantifierAgent* with 50 seeds and plot the change (Δ) in the sum of mean coefficient of variation across all criteria against the number of seeds, in Figure 6. For each criterion, we compute the absolute difference with the mean coefficient of variation calculated when using $n - 1$ seeds, summing up the absolute differences across all criteria. According to the plot, after approximately 18 seeds, the magnitude of mean coefficient of variation stabilizes and becomes rather trivial. In almost all cases, the mean coefficient of variation is around or below 0.5, which is relatively small, suggesting that *QuantifierAgent* is quite robust.

6.2.2 Discriminative Power

It is crucial to ensure the quality of quantification of each criterion. Ideally, this validation would involve comparisons with known pairwise samples,

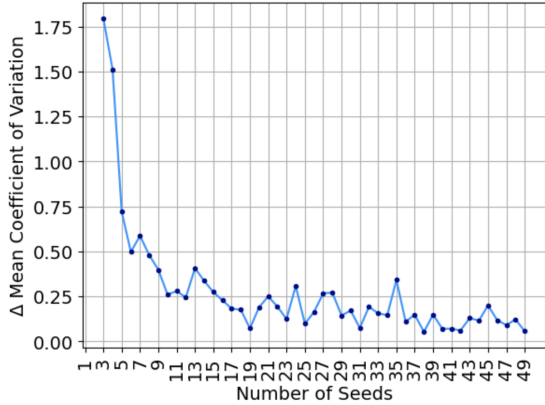


Figure 6: Δ sum of mean coefficient of variation across all criteria with increasing number of seeds.

where sample S^+ is definitively superior to S^- for a given criterion. If the evaluator also confirms superiority of S^+ w.r.t S^- , it has robust quantification. However, due to rapid expansion of LLM-powered applications, obtaining annotated data for many tasks is often unfeasible. Therefore, we propose using synthetically altered versions of samples for verification. Let us assume we have an alternative disturbed version of sample S , which is called S' . Assuming sample S is more likely to outperform its disturbed version S' , our assessment should confirm this assumption by assigning better quantified performance S in comparison to S' . In experiments with mathematical problems, we introduced random noise by removing portions of the solution sentences from AutoGen, VanillaSolver, and ReAct’s results respectively, expecting that criteria like ‘Completeness’ or ‘Clarity’ would show be higher in S than in S' . We disturbed solutions by removing 25% of the sentences and assessed the *QuantifierAgent*’s performance. As shown in Fig. 7, criteria measuring aspects like ‘Clarity’ and ‘Completeness’ were lower in disturbed solutions (lighter bars), confirming *QuantifierAgent*’s high discriminative power and effectiveness.

We have already filtered out the criteria that were unstable, i.e., those that had a high mean standard deviation and dispersion when being quantified in the previous section. We report the results of the *QuantifierAgent* quantifying differences between original and disturbed samples on the comprehensive set of criteria shown in Appendix, as shown in Fig. 13 for the math problem-solving. In most cases, the *QuantifierAgent* quantifies the disturbed output to be worse than the original task output. We believe analyzing the *QuantifierAgent*’s performance will enhance the reliability, trustworthiness,

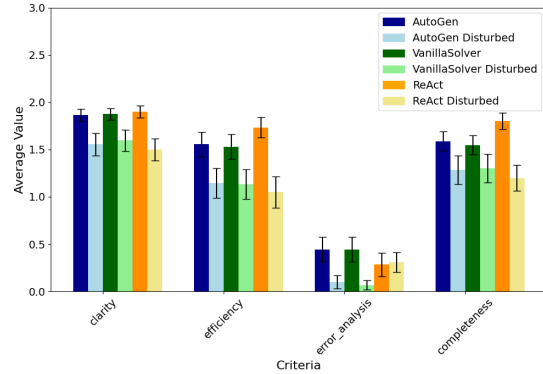


Figure 7: Assessment of original and disturbed solutions on Math dataset (discriminative power study).

and explainability in evaluations..

6.2.3 VerifierAgent

After modifying the list of criteria (Sec. 6.2.1), we have developed a stable and robust list of criteria that the *QuantifierAgent* can reliably quantify. Further, we also proposed a method for assessing whether the criteria can distinguish between noise-adversarially attacked samples and the original ones. These two tests will serve as input for the *VerifierAgent* (described in Algorithm 1), which can also have its threshold tuned for different applications. For instance, one might prioritize the stability of the criteria, while another may value the discriminative power of the *AgentEval* for specific applications. As such, the *VerifierAgent* will modify and update the criteria based on to what extend they pass the two tests, i.e., if the mean coefficient of variation is below a specific threshold and the percentage of adversarial testing it has passed. The *VerifierAgent* will then update the criteria if necessary. We believe that having a *VerifierAgent* would help continuously updating the criteria as needed because, by improving the systems, we may require new criteria that were not previously necessary for utility assessment.

7 Conclusions and Future Work

We introduced the *AgentEval* framework, designed to swiftly gauge the utility of arbitrary LLM-powered agentic applications. Our framework leverages recent findings suggesting LLMs as a scalable and cost-effective alternative to human evaluations for open-ended tasks. *AgentEval* consists of three agents: *CriticAgent* suggests criteria based on task descriptions and executions of the applications, *QuantifierAgent* quantifies how well the application flow aligns with these crite-

ria, and *VerifierAgent* modifies the list of criteria if needed. This framework is customizable, adaptable, and can operate in various modes, employing combinations of LLMs, human inputs, and tools. We believe that suggested *AgentEval*'s utility extends beyond immediate performance. It can uncover new system capabilities over time and adapt to changes in user needs tracked by developers. *AgentEval* can also enable developers to assess the alignment between application behavior and suggested user requirements, providing them with insights into areas for improvement. In summary, our contributions include introducing the *AgentEval* framework, and conducting a robust analysis of its performance across various datasets and baselines. *AgentEval* represents a significant step towards assessing LLM-powered applications.

8 Limitations and Ethics

8.1 Limitations

Here, we discuss some limitations of the *AgentEval* framework. Firstly, the performance of the *AgentEval* is highly dependent on the quality of the output logs of the applications. Flaws or limitations in these outputs can significantly impact the framework's ability to accurately assess utility.

Secondly, our experiments were conducted exclusively with closed-source LLMs, specifically with GPT-4. This may limit the generalizability of our findings. Plans to include a broader array of LLMs, including open-source models, are considered for future studies to validate and possibly enhance the robustness of our conclusions. Additionally, the tests conducted were limited to specific scenarios within math problem solving and household tasks. Expanding the diversity of test scenarios could help in understanding the broader applicability of the framework.

Thirdly, while *AgentEval* employs a novel methodology leveraging LLMs to estimate utility, the absence of human evaluation in our validation process could be viewed as a drawback. Human evaluations provide unique insights, especially in subjective aspects of utility that automated systems might overlook. However, such evaluations are often cost-prohibitive and logistically challenging, restricting our ability to implement them within this study. Especially do developers of agentic LLM-powered applications who needs insights fast as they go with the deployments.

Lastly, as LLM technologies evolve, the criteria

and metrics used for evaluation may need to be updated or revised. What works for assessing current LLMs may not hold as these models become more advanced. Continuous updates to the evaluation framework will be necessary to keep pace with technological advancements.

8.2 Ethics

To the best of our knowledge, we did not violate any code of ethics with the experiments done in this paper. We reported technical details and results, with details in the main paper, Appendix, and code release. Our experimental results are an outcome of a Machine Learning model.

Our *AgentEval* system has a variety of uses in real world settings, such as improving applications for end users or helping developers. However, we caution that it must be used carefully, as the outputs are from a ML model and can have real world consequences, if used incorrectly.

These and many other related issues are important aspects to consider when deploying a system like *AgentEval* in the real world.

Acknowledgement

We thank Besmira Nushi, Victor Dibia, and Adam Fourney for inspiring discussions for earlier versions of the *AgentEval* framework.

References

- Ali Ahmadvand, Negar Arabzadeh, Julia Kiseleva, Patricio Figueroa Sanz, Xin Deng, Sujay Jauhar, Michael Gamon, Eugene Agichtein, Ned Friend, et al. 2022. Supporting complex information-seeking tasks with implicit constraints. *arXiv preprint arXiv:2205.00584*.
- Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. 2024. Generative information retrieval evaluation. *arXiv preprint arXiv:2404.08137*.
- Albert Andrew. 2024. Potential applications and implications of large language models in primary care. *Family Medicine and Community Health*, 12(Suppl 1).
- Negar Arabzadeh, Ali Ahmadvand, Julia Kiseleva, Yang Liu, Ahmed Hassan Awadallah, Ming Zhong, and Milad Shokouhi. 2022. Preme: Preference-based meeting exploration through an interactive questionnaire. *arXiv preprint arXiv:2205.02370*.
- Negar Arabzadeh, Amin Bigdeli, and Charles LA Clarke. 2024a. Adapting standard retrieval benchmarks to evaluate generated answers. In *European*

- Conference on Information Retrieval*, pages 399–414. Springer.
- Negar Arabzadeh, Amin Bigdeli, Radin Hamidi Rad, and Ebrahim Bagheri. 2023. Quantifying ranker coverage of different query subspaces. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2298–2302.
- Negar Arabzadeh and Charles LA Clarke. 2024. A comparison of methods for evaluating generative ir. *arXiv preprint arXiv:2404.04044*.
- Negar Arabzadeh, Julia Kiseleva, Qingyun Wu, Chi Wang, Ahmed Awadallah, Victor Dibia, Adam Fournery, and Charles Clarke. 2024b. Towards better human-agent alignment: Assessing task utility in llm-powered applications. *arXiv preprint arXiv:2402.09015*.
- Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the utility of search engine result pages: an information foraging based measure. In *The 41st International ACM SIGIR conference on research & development in information retrieval*, pages 605–614.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. [A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity](#).
- Muneera Bano, Didar Zowghi, and Jon Whittle. 2023. Exploring qualitative research using llms. *arXiv preprint arXiv:2306.13298*.
- Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.
- Alexey Borisov, Julia Kiseleva, Ilya Markov, and Maarten de Rijke. 2018. Calibration: A simple way to improve click models. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1503–1506.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.
- Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate. *arXiv preprint arXiv:2401.16788*.
- Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.
- Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. 2019. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*, pages 41–75. Springer.
- Victor Dibia, Adam Fournery, Gagan Bansal, Forough Poursabzi-Sangdeh, Han Liu, and Saleema Amershi. 2023. [Aligning offline metrics and human judgments of value for code generation models](#).
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.
- Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.
- Siqing Huo, Negar Arabzadeh, and Charles LA Clarke. 2023. Retrieving supporting evidence for llms generated answers. *arXiv preprint arXiv:2306.13781*.
- Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023.

- Multi-dimensional evaluation of text summarization with in-context learning. *arXiv preprint arXiv:2306.01200*.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.
- Julia Kiseleva, Eric Crestan, Riccardo Brigo, and Roland Dittel. 2014. Modelling and detecting changes in user satisfaction. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1449–1458.
- Julia Kiseleva and Maarten de Rijke. 2017. Evaluating personal assistants on mobile devices. *arXiv preprint arXiv:1706.04524*.
- Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, Arthur Szlam, Yuxuan Sun, Katja Hofmann, Marc-Alexandre Côté, Ahmed Awadallah, Linar Abdrazakov, Igor Churin, Putra Manggala, Kata Naszadi, Michiel van der Meer, and Taewoon Kim. 2022a. [Interactive grounded language understanding in a collaborative environment: Iglu 2021](#). In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 146–161. PMLR.
- Julia Kiseleva, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Côté, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail Burtsev, Maartje ter Hoeve, Zoya Volovikova, Aleksandr Panov, Yuxuan Sun, Kavya Srinet, Arthur Szlam, Ahmed Awadallah, Seungeun Rho, Taehwan Kwon, Daniel Wontae Nam, Felipe Bivort Haiek, Edwin Zhang, Linar Abdrazakov, Guo Qingyam, Jason Zhang, and Zhibin Guo. 2022b. [Interactive grounded language understanding in a collaborative environment: Retrospective on iglu 2022 competition](#). In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 204–216. PMLR.
- Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anatasakos. 2016a. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 45–54.
- Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anatasakos. 2016b. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 121–130.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for "mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.
- Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2023b. Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation. *arXiv preprint arXiv:2310.19740*.
- Ziming Li, Julia Kiseleva, Alekh Agarwal, Maarten de Rijke, and Ryen W White. 2020. Optimizing interactive systems via data-driven objectives. *arXiv preprint arXiv:2006.12999*.
- Ziming Li, Dookun Park, Julia Kiseleva, Young-Bum Kim, and Sungjin Lee. 2021. Deus: A data-driven approach to estimate user satisfaction in multi-turn dialogues. *arXiv preprint arXiv:2103.01287*.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023a. [Holistic evaluation of language models](#).
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023b. [Encouraging divergent thinking in large language models through multi-agent debate](#).
- Alex Liu and Min Sun. 2023. From voices to validity: Leveraging large language models (llms) for textual analysis of policy stakeholder interviews. *arXiv preprint arXiv:2312.01202*.
- Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, and Li Zhang. 2024. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*.

- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023. Agent-bench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.
- Yanchen Liu, Timo Schick, and Hinrich Schütze. 2022. Semantic-oriented unlabeled priming for large-scale language models. *arXiv preprint arXiv:2202.06133*.
- Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Query performance prediction using relevance judgments generated by large language models. *arXiv preprint arXiv:2404.01012*.
- Grégoire Mialon, Clémentine Fourier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*.
- Shrestha Mohanty, Negar Arabzadeh, Julia Kiseleva, Artem Zholus, Milagro Teruel, Ahmed Awadallah, Yuxuan Sun, Kavya Srinet, and Arthur Szlam. 2023. Transforming human-centered ai collaboration: Redefining embodied agents capabilities through interactive grounded language instructions. *arXiv preprint arXiv:2305.10783*.
- Shrestha Mohanty, Negar Arabzadeh, Andrea Tupini, Yuxuan Sun, Alexey Skrynnik, Artem Zholus, Marc-Alexandre Côté, and Julia Kiseleva. 2024. Idat: A multi-modal dataset and toolkit for building and evaluating interactive task-solving agents. *arXiv preprint arXiv:2407.08898*.
- Vivek Myers, Andre Wang He, Kuan Fang, Homer Rich Walke, Philippe Hansen-Estruch, Ching-An Cheng, Mihai Jalobeanu, Andrey Kolobov, Anca Dragan, and Sergey Levine. 2023. [Goal representations for instruction following: A semi-supervised language interface to control](#). In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 3894–3908. PMLR.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.
- Jiayi Pan, Yichi Zhang, Nicholas Tomlin, Yifei Zhou, Sergey Levine, and Alane Suhr. 2024. Autonomous evaluation and refinement of digital agents. *arXiv preprint arXiv:2404.06474*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.
- Mahsa Seifkar, Linh Nhi Phan Minh, Negar Arabzadeh, Charles LA Clarke, and Mark D Smucker. 2023. A preference judgment tool for authoritative assessment. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3100–3104.
- Anna Sepiarskaia, Julia Kiseleva, Filip Radlinski, and Maarten de Rijke. 2018. Preference elicitation as an optimization problem. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 172–180.
- Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2019. [ALFRED: A benchmark for interpreting grounded instructions for everyday tasks](#). *CoRR*, abs/1912.01734.
- Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020a. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2020b. Alworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.
- Bernd Carsten Stahl and Damian Eke. 2024. The ethics of chatgpt—exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74:102700.
- Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.
- Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2023. Do llms exhibit human-like response biases? a case study in survey design. *arXiv preprint arXiv:2311.04076*.
- Teemu Vahtola, Mathias Creutz, and Jörg Tiedemann. 2022. It is not easy to detect paraphrases: Analysing semantic similarity with antonyms and negation using the new semantoneg benchmark. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 249–262.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11.
- Jindong Wang, Xixu HU, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, Binxing Jiao, Yue

- Zhang, and Xing Xie. 2023b. [On the robustness of chatGPT: An adversarial and out-of-distribution perspective](#). In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.
- Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023c. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.
- Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016a. Detecting good abandonment in mobile search. In *Proceedings of the 25th International Conference on World Wide Web*, pages 495–505.
- Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016b. Is this your final answer? evaluating the effect of answers on good abandonment in mobile search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 889–892.
- Kyle Williams and Imed Zitouni. 2017. Does that mean you’re happy? rnn-based modeling of user interaction sequences to detect good abandonment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 727–736.
- Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Auto-gen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- Chi Zhang, Penglin Cai, Yuhui Fu, Haoqi Yuan, and Zongqing Lu. 2023. Creative agents: Empowering agents with imagination for creative tasks. *arXiv preprint arXiv:2312.02519*.
- Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. [Through the lens of core competency: Survey on evaluation of large language models](#). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109, Harbin, China. Chinese Information Processing Society of China.

A Appendix

A.1 Task Examples

In Fig. 8 and 9, we display examples of Math problems and ALFWorld house-holding tasks solved with AutoGen.

A.2 Math Problem Solving Benchmark

For math problem solving, although success is clearly defined, multiple solutions exist for accomplishing the objectives. The MATH dataset, originally is a substantial collection of 12,500 challenging mathematics problems from high school competitions (Hendrycks et al., 2021b). Each problem comes with a step-by-step solution, enabling models to learn how to generate both derivations and explanations. The dataset covers a wide range of mathematical subjects and is tagged by difficulty levels, offering a nuanced measure of model performance across various aspects of mathematical problem-solving.

This dataset is particularly suitable for testing multi-agent systems for several reason including: (i) The problems in the MATH dataset are not simple computations but require a deep understanding of mathematical concepts, heuristics, and problem-solving strategies. (ii) Since the dataset includes step-by-step solutions, it allows for the assessment of an agent’s ability to learn and reason through a problem, not just its ability to arrive at the correct answer. (iii) The variety of subjects and difficulty levels in the MATH dataset enables a comprehensive evaluation of a system’s versatility and adaptability in different mathematical domains which is crucial for multi-agent systems that are expected to operate across a range of scenarios.

Similar to math problem experimental setup in Wu et al. (2023), we carry out two experimental evaluations which involves 120 problems from the most challenging category, and includes 20 problems each from six different categories, of number theory, counting and probability, prealgebra, algebra, intermediate algebra, and precalculus.

A.3 ALFWorld House-holding Task

A.3.1 ALFWorld Dataset

ALFWorld, presents a set of language-based interactive decision-making tasks within simulated household environments (Shridhar et al., 2020b). This benchmark is distinguished by its diversity of tasks, offering a comprehensive platform

Table 2: Verification Criteria for ALFWorld Household-ing Tasks.

Criteria	Description	Accepted Values
Task Understanding	How well the participant was able to comprehend the problem set and follow the task instructions	- Excellent (4) - Good (3) - Average (2) - Poor (1) - Terrible (0)
Plan Making	The ability of the participant to strategize and make a plan for tackling the task.	- Excellent (4) - Good (3) - Average (2) - Poor (1) - Terrible (0)
Action Decision	The participant’s decision-making skills in choosing the right action to perform.	- Excellent (4) - Good (3) - Average (2) - Poor (1) - Terrible (0)
Action Execution	How effectively the participant is able to execute the chosen action.	- Excellent (4) - Good (3) - Average (2) - Poor (1) - Terrible (0)
Response to Feedback	How well the participant adapts his/her next steps based on the feedback from the environment	- Excellent (4) - Good (3) - Average (2) - Poor (1) - Terrible (0)
Correctness of Action	The correctness of the action performed by the participant with respect to the available actions and the current context	- Correct (1) - Incorrect (0)
Use of Terminate	Whether the participant uses the ‘TERMINATE’ command appropriately	- Appropriate (1) - Inappropriate (0)

for testing AI and multi-agent systems. This benchmark is particularly suited for such evaluations because first, ALFWorld is the first interactive parallel environment that aligns text descriptions and commands with physically embodied robotic simulation. It extends two prior works: TextWorld, an engine for interactive text-based games, and ALFRED, a large-scale dataset for vision-language instruction following in embodied environments (Shridhar et al., 2020a; Côté et al., 2019). The cross-modality framework of this benchmark allows for a variety of embodied tasks with corresponding text-based counterparts, allowing agents to be trained and evaluated in both the language and embodied world. In addition, ALFWorld supports the development of agents that can reason both abstractly and execute actions concretely, mimicking human-like decision-making processes in varying contexts. Finally, the dataset’s inclusion of a wide range of tasks, from household chores to more intricate problem-solving scenarios, provides a comprehensive testbed for evaluating the adaptability and problem-solving capabilities of AI and multi-agent systems. In general, the dataset allows agents to explore, interact, and learn in an abstract language environment before dealing with the complexities of an embodied environment.

```

"problem": "How many positive and negative integers is 12 a multiple of?"
"level": "Level 5"
"type": "Prealgebra"
"response_with_ans":
"messages": [
{
"content": "How many positive and negative integers is 12 a multiple of?",
"role": "user"
},
{
"content": "To determine how many positive and negative integers 12 is a multiple of, we need to find the divisors of 12. Then we will double the number of divisors, as each positive divisor has a corresponding negative divisor. Let's use a Python code snippet to find all the divisors of 12 and count them.",
"role": "assistant"
}
],
"time": 17.360238790512085

```

```

1 num = 12
2 divisors = [i for i in range(1, num + 1) if num % i == 0]
3 positive_divisors_count = len(divisors)
4 negative_divisors_count = positive_divisors_count
5 total_divisors_count = positive_divisors_count +
    ↪ negative_divisors_count
6 print(f"Positive divisors of 12: {divisors}")
7 divisors_negative = [-x for x in divisors]
8 print(f"Negative divisors of 12: {divisors_negative}")
9 print(f"Total divisors count: {total_divisors_count}")

```

Figure 8: An example of Math Problem solving sample with solution generated by AutoGen

A.3.2 Solutions for ALFWorld benchmark

As for the solutions to assess for ALFWorld Household tasks, similar to (Wu et al., 2023), we consider ReAct (Yao et al., 2022) as well as AutoGen with two agents and AutoGen with three agents (Wu et al., 2023). ReAct is an agent that operates within the ALFWorld environments and is responsible for suggesting plans and executing actions. On the other hand, AutoGen Two-Agent System consists of an LLM-backed assistant agent responsible for suggesting plans, and an executor agent responsible for executing actions in the ALFWorld environments. Both ReAct and this solution occasionally struggles with leveraging basic commonsense knowledge about the physical world,

which can lead to repetitive errors and getting stuck in loops. In AutoGen with three agents, a grounding agent is provided just for the sake of critical common sense knowledge whenever the system exhibits early signs of recurring errors.

A.3.3 AgentEval Results for ALFWorld

To study the generalizability of AgentEval, we repeat the experiments in 5.2 for ALFWorld, in which real-world household environments are emulated through textual interfaces (Shridhar et al., 2020b). We provide the criteria created for this task as well as the results for three solutions of this task in Tab. 2 and Fig. 10, respectively. Following the extraction of a set of criteria as detailed in Tab. 2, these criteria are passed to the *QuantifierAgent* for quan-

```

{
  "content": "Perform actions and interact with a household to solve a task. At the beginning of
  ↪ your interactions, you will be given the detailed description of the current environment and
  ↪ your goal to accomplish. For each of your turn, you should choose from two actions: "think"
  ↪ or "Action". If you choose "think", you should first think about the current condition
  ↪ and plan for your future actions, and then output your action in this turn. Your output must
  ↪ strictly follow this format:"think: your thoughts."
  "Action: your next action\n"; If you choose "ACTION", you should directly output the action in
  ↪ this turn. Your output must strictly follow this format:"ACTION: your next action".
  After each of your turn, the environment will give you immediate feedback based on which you
  ↪ should plan your next few steps. if the environment output "Nothing happened", that means the
  ↪ previous action is invalid and you should try more options.
  Reminder:
  1. the action must be chosen from the given available actions. Any actions except provided
  ↪ available actions will be regarded as illegal.
  2. Take the target household immediatly after you find it.
  3. Reply 'TERMINATE' only under two circumstances: a). The task has given you clear instructions
  ↪ to return 'TERMINATE' b). The task seems insolvable.\n Here are two examples.\nYou are in
  ↪ the middle of a room. Looking quickly around you, you see a cabinet 13, a cabinet 12, a
  ↪ cabinet 11, a cabinet 10, a cabinet 9, a cabinet 8, a cabinet 7, a cabinet 6, a cabinet 5, a
  ↪ cabinet 4, a cabinet 3, a cabinet 2, a cabinet 1, a coffeemachine 1, a countertop 1, a
  ↪ diningtable 1, a drawer 1, a fridge 1, a garbagecan 1, a microwave 1, a shelf 3, a shelf 2, a
  ↪ shelf 1, a sinkbasin 1, a stoveburner 4, a stoveburner 3, a stoveburner 2, a stoveburner 1,
  ↪ and a toaster 1.\nYour task is to: put a hot apple in fridge.\n",
    "role": "user"},
  {
    "content": "THOUGHT: The task is to put a hot apple in fridge. So i need to first find an
    ↪ apple and heat it. First I need to find an apple. An apple is more likely to appear in
    ↪ diningtable (1), coffeetable (1), fridge (1), drawer (1), cabinet (1-13), garbagecan (1). I
    ↪ can check one by one, starting with fridge 1.\n ACTION: go to diningtable 1\n",
    "role": "assistant"},
  {
    "content": "On the diningtable 1, you see a apple 1, a bread 1, a fork 3, a fork 2, a fork 1,
    ↪ a mug 2, a peppershaker 3, a plate 2, a pot 1, a soapbottle 3, a spatula 1, a spoon 3, a
    ↪ spoon 2, a spoon 1, a winebottle 3, and a winebottle 2.\n",
    "role": "user"},
  {
    "content": "ACTION: take apple 1 from diningtable 1\n",
    "role": "assistant"},
  {
    "content": "You pick up the apple 1 from the diningtable 1.\n",
    "role": "user"},
  {
    "content": "THOUGHT: I noticed that there is a microwave 1 in the room, i will use it to heat
    ↪ the apple.\n ACTION: go to microwave 1\n",
    "role": "assistant"},
    ...
    ...
  {
    "content": "You open the microwave 1. The microwave 1 is open. In it, you see a egg 1.",
    "role": "user"
  },
  {
    "content": "ACTION: heat mug 2 with microwave 1",
    "role": "assistant"},
  {
    "content": "You heat the mug 2 using the microwave 1.Task success, now reply TERMINATE\n",
    "role": "user"},
  {
    "content": "TERMINATE",
    "role": "assistant"},
  {
    "content": "TERMINATE",
    "role": "user"}
}

```

Figure 9: An example of ALFWorld Householding task from “Pick, Heat, then Place” category.

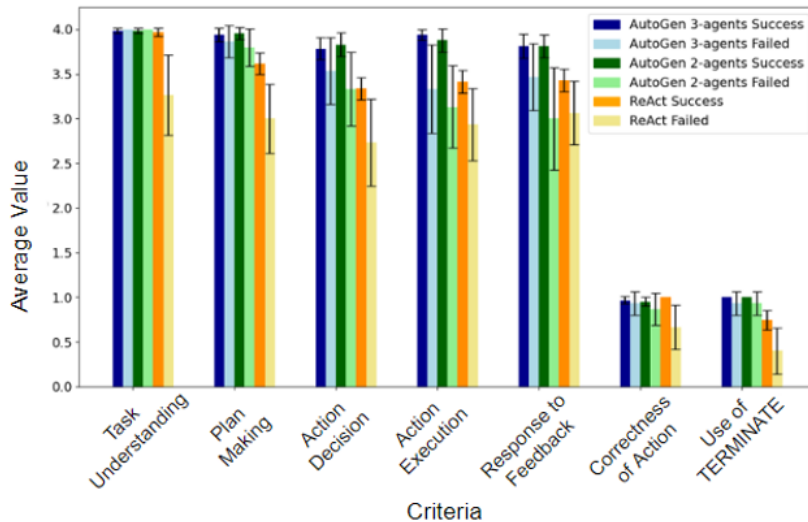


Figure 10: AgentEval assessment of three different solutions on ALFWorld house-holding Tasks categorized by success and failed cases.

tification on each sample. Figure 10 presents the results for three introduced solutions: AutoGen with two agents, AutoGen with three agents, and ReAct, on the 134-test set from (Wu et al., 2023). Further, we also verify the measured performance of AgentEval, by comparing the failed/success cases in the right plot of Figure 10. As confirmed here and in math problem solving, AgentEval is clearly able to distinguish between different solutions in terms of different criterion.

We provide an example of AgentEval applied to the ALFWorld house-holding task, as mentioned in Sec. 5.2 in which real-world household environments is emulated through textual interfaces (Shridhar et al., 2020b). When running the *CriticAgent* on this task, it identified specific criteria such as “Task understanding”, “Plan making” and “Response to Feedback” as outlined in Tab. 2. We consulted researchers deeply involved with these tasks, and their expertise confirmed that these criteria are critically relevant and significant similar to (Li et al., 2023b). For example, given that these tasks are language-based and require interactive decision-making, an agent in ALFWorld is tasked with high-level objectives, such as placing a hot apple in the fridge, and must navigate and interact with a simulated household environment to achieve these objectives. Therefore, criteria displayed in Tab. 2 satisfy the assessment of this task. While the criteria are pretty self-descriptive, about the criterion “Use of TERMINATE” we note that the agent is prompted to use the term “TERMINATE” upon task completion, which is closely correlated

with task success.

Following the extraction of a set of criteria as detailed in Tab 2, these criteria are passed to the *QuantifierAgent* for quantification on each sample. Figure 10 presents the results for three introduced solutions: AutoGen with 2 agents, AutoGen with 3 agents, and ReAct, on the 134-test set from (Wu et al., 2023). It is important to note that all criteria, except “Use of TERMINATE” and “Correctness of Action” employ a five-level grading system, while these two criteria are binary. From this figure, it is evident that ReACT performs notably worse across all criteria, while AutoGen with 2 agents and 3 agents demonstrate competitive performance. We also categorizes the 134 games into groups of failed and successful ones. Similar to Fig. 3, darker colors represent performance in successful cases for each solution, while lighter colors represent performance in failed cases. AutoGen 3-agent, AutoGen 2-agent, and ReAct are represented by blue, green, and orange, respectively. For most criteria, the distinction between failed and successful cases is clear, even within a 95% confidence interval. However, for certain criteria, such as “Task understanding” all solutions, whether they failed or succeeded, exhibit very similar performance. This could be interpreted as either (1) all solutions have a good understanding of the task, even if they fail to complete it, (2) this criterion may be redundant, as it does not provide additional information among these three solutions or (3) the *QuantifierAgent* is unable to score the criterion in a meaningful way. We refrain from concluding which criteria are most

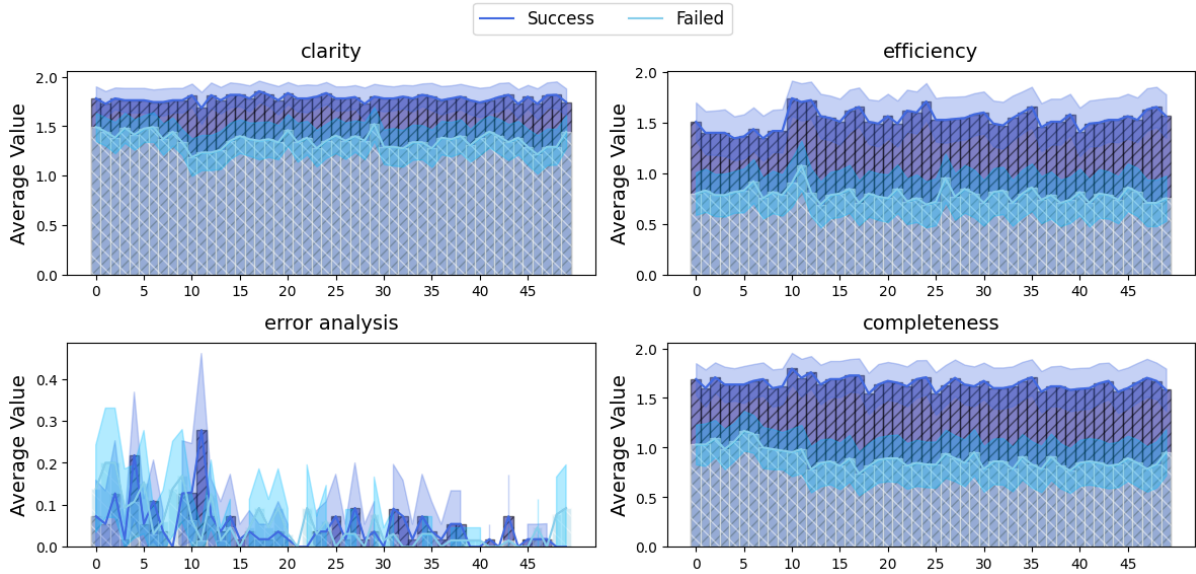


Figure 11: Quantifier Robustness on criteria of Math Problem Solving problem. Each bar represent the average performance of success (dark blue “//”) and failed (light blue “\”) cases and 95% interval on each set is shaded across the average point. The two plots are overlaid.

Table 3: Example pairs of similar criteria.

- Problem Difficulty: The complexity of the math problem that has been solved.
- Problem Complexity: The level of difficulty of the problem.
- Innovativeness: The novelty and creativity in the approach to solve the problem
- Innovation: The ability to solve a problem using a unique or creative method not commonly known.
- Time Taken: The time taken to solve the problem.
- Time to Completion: The amount of time taken to solve the problem completely
- Understandability: The clarity and ease of comprehension of the solution provided.
- Readability: How easy it is to comprehend the provided solution.

suitable for this specific task. Instead, we emphasize the importance of conducting a more in-depth analysis of performance beyond success rates, tailored to one’s goals and application requirements. Later, we show that how using *VerifierAgent* could be helpful in identifying criteria with higher discriminative power and more robustness.

A.4 Robustness Analysis

A.4.1 Similar Criteria

As explained in Section 6.1, there might be cases where some criteria are pointing to the same concepts with different wordings. In these cases, we need to merge the similar criteria to avoid having redundant criteria. Table 3 shows some of these examples.

A.4.2 Quantifier Robustness

To study the robustness of the *QuantifierAgent*, we selected a specific subset of criteria related to mathematical problems, as detailed in Table 1, and conducted 50 runs of the quantifier agent on the 120 problems described in Section 4.1. Our expectation is to observe consistent quantified performance for each of the criteria. In Fig. 11, we present the distribution of quantified performance across 50 runs for both successful and failed cases, focusing on the five selected criteria. A consistently horizontal performance trend indicates greater robustness in the quantifier, whereas more fluctuations in the figure suggest less robustness and a noisier performance of the agent.

As shown in the results, for four out of the five generated criteria, we consistently observe steady performance. Not only do the success cases consistently outperform the failed cases, but their performance also falls within a similar range across runs. However, when it comes to the “error analysis” criterion, we observe a more variable performance of the quantifier. It does not consistently predict one group (success or failed) to perform better than the other, and the quantifier’s performance varies across different runs. This suggests that the *AgentEval* tool may not exhibit promising robustness for this particular criterion. The underlying issues could be either the criterion itself lacks clarity and appropriateness for the task, or the *QuantifierA-*

gent struggles to quantify this criterion effectively. In either case, it is advisable to either modify or eliminate this criterion to enhance trustworthiness and reliability. We further show that *VerifierAgent* is designed to take care of such criteria.

We recognize the importance of thoroughly investigating the robustness of each criterion in quantification studies. This analysis is crucial as it sheds light on the stability of each criterion. Moreover, when ground truths are available, such as in cases of success versus failure, they provide a benchmark to validate our assessments. Additionally, it is important to acknowledge that not all criteria exhibit the same level of robustness. This variability demands careful consideration during evaluations, especially given the non-deterministic nature of LLMs. Such awareness is essential to ensure the reliability and accuracy of our assessments in the dynamic field of LLMs.

A.5 VerifierAgent

Algorithm 1 shows how *VerifierAgent* works. To make *VerifierAgent* works, we need to study the stability of proposed criteria as well as how robust they are w.r.t the injected noise.

A.5.1 Criteria Robustness

we first report the full criteria list for Math problems solving and ALFWorld household tasks when running the *CriticAgent* and *QuantifierAgent* for 50 times after consolidation (as described in section 6.1) in Tab 4 and 5. This process would exclude criteria that have mean standard deviation above a certain threshold and criteria that have a higher or equivalent average score for adversarial task output than the original task output. This does not necessarily mean these criteria are bad criteria, but rather suggests the *QuantifierAgent* may not be able to reliably quantify these criteria and thus it might be better to exclude them from the final score assigned to a sample. As such, similar to Fig. 6, we report the mean of coefficient variation for ALFWorld task in Fig. 12. We note that having almost all of the coefficient below 0.5 indicate high level of robustness of *QuantifierAgent* on the verified set of criteria by *VerifierAgent* on AlfWorld dataset.

A.5.2 Adversarial Attacks

We construct adversarial samples by randomly dropping a portion of sentences in the LLM assistant’s response from the original task output. We verify the *QuantifierAgent* against the adversarial

Algorithm 1 *VerifierAgent*

```

1: for  $i = 1, 2, \dots, 50$  do
2:   Run CriticAgent with  $seed = i$  to obtain a
   set of criteria  $C_i$ 
3: end for
4: Obtain summarized_criteria by using another
   LLM agent to summarize  $C_1, C_2, \dots, C_{50}$ .
5: for  $i = 1, 2, \dots, 18$  do
6:   for all  $s$  in  $S$  do
7:     Run QuantifierAgent with  $seed = i$  on
     sample  $s$ 
8:   end for
9: end for
10: for all crit in summarized_criteria do
11:   for all  $s$  in  $S$  do
12:     Compute the coefficient of variation of
      $s$ 's quantified result with respect to crit across
     all  $seed$ 
13:   end for
14:   Compute mean coefficient of variation by
     averaging all sample's coefficient of variation
15: end for
16: final_criteria  $\leftarrow \square$ 
17: for all crit in summarized_criteria do
18:   if crit has a mean coefficient of variation
     within a certain range, and crit has decent ad-
     versarial testing performance then
19:     Add crit to final_criteria
20:   end if
21: end for
22: To evaluate future tasks, use final_criteria with
   QuantifierAgent.

```

samples. We used three different benchmarks for adversarial testing, namely AutoGen, ReAct and Vanilla Solver. As shown in Fig. 13 for the ALFWorld dataset), in most cases the *QuantifierAgent* quantifies the adversarial task output to be worse off than the original task output. We believe that such an analysis of the quantifier agent’s performance will yield valuable insights for enhancing reliability, trustworthiness, and explainability in performance evaluation.

One interesting observation here is that there maybe interdependence among some criteria. For example *level appropriateness* is defined as "*How well-suited the solution provided by the system is for the given problem's level*", which is dependent on the criterion *problem level*. This observation gives insight into potential future improvements to the current pipeline. We may first extract some

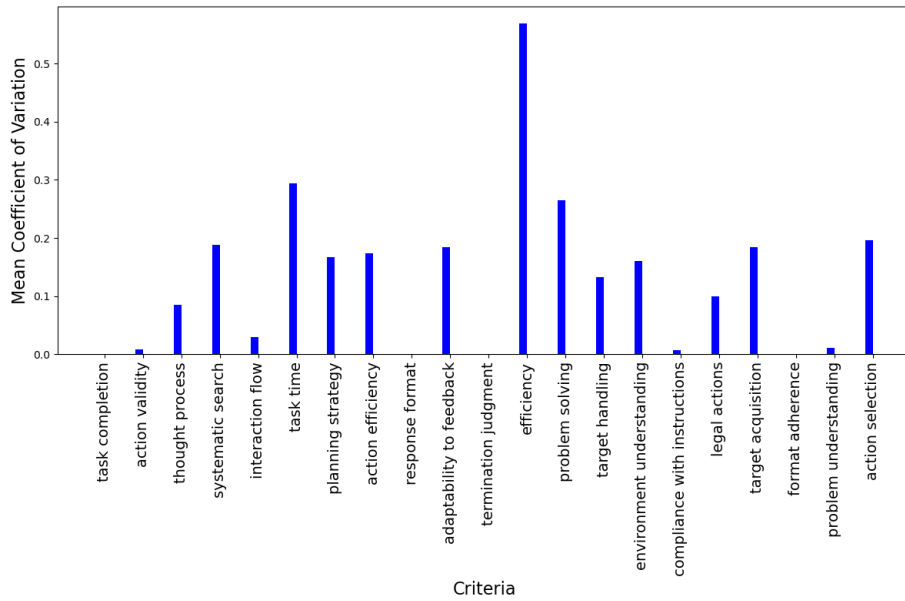


Figure 12: Evaluating the *QuantifierAgent*'s robustness on ALFWorld dataset: the mean coefficient of variation of quantified results across $n = 18$ seeds.

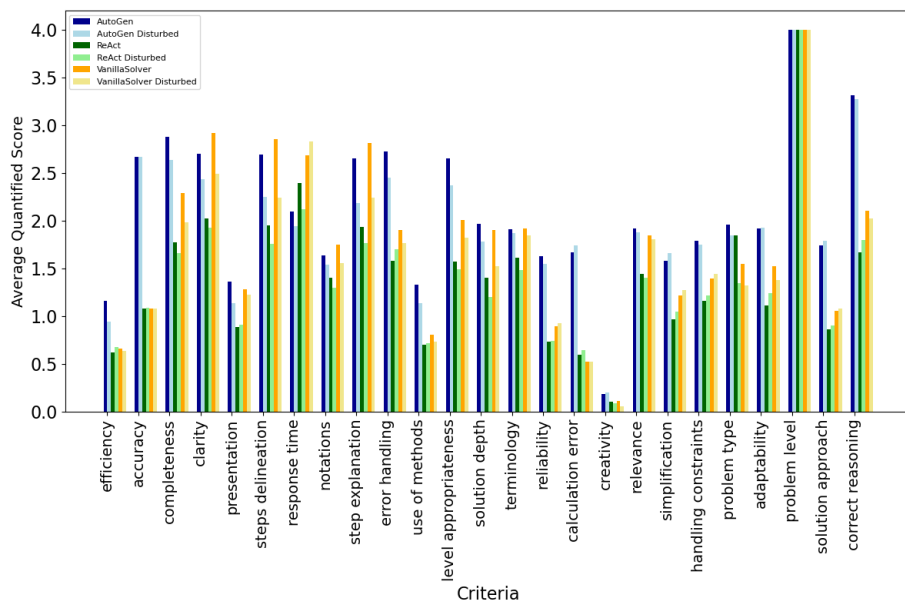


Figure 13: *QuantifierAgent* Verification on original set of task solutions against the disturbed task solutions on Math Problem Solving dataset.

characteristics of the task output, such as categorical criteria like *problem type* and *problem level*, and then potentially generate different criteria and quantify the task output differently based on these characteristics.

Table 4: Comprehensive Verification Criteria for Math-Problems.

Criteria	Description	Accepted Values
efficiency	The conciseness of the solution and the use of the most efficient method to solve the problem.	– highly_efficient (2) – moderately_efficient (1) – inefficient (0)
accuracy	The correctness of the solution provided for the math problem.	– 100% - Completely correct (4) – 75% - Almost correct (3) – 50% - Mostly correct (2) – 25% - Partially correct (1) – 0% - Completely incorrect (0)
completeness	The extent to which the solution covers all aspects of the problem.	– 100% - Fully complete (4) – 75% - Almost complete (3) – 50% - Mostly complete (2) – 25% - Partially complete (1) – 0% - Not complete* (0)
clarity	The ease with which the solution can be understood by the target audience.	– 100% - Very clear (4) – 75% - Mostly clear (3) – 50% - Fairly clear (2) – 25% - Somewhat clear (1) – 0% - Not clear (0)
presentation	The organization and presentation of the solution, including proper use of notation, symbols, and formatting.	– excellent (2) – fair (1) – poor (0)
steps delineation	How well the solution breaks down the problem-solving process into smaller, manageable steps.	– 100% - All steps delineated (4) – 75% - Most steps delineated (3) – 50% - Some steps delineated (2) – 25% - Few steps delineated (1) – 0% - No steps delineated (0)
response time	The time taken to provide the solution	– >5 min (5) – 3-5 min (4) – 1-3 min (3) – 31-60 sec (2) – 16-30 sec (1) – 0-15 sec (0)
notations	The notations used in the problem solution are appropriate and consistent.	– consistent (2) – mostly consistent (1) – inconsistent (0)
steps explanation	The extent to which each step in the solution is explained.	– all steps (4) – most steps (3) – half steps (2) – some steps (1) – none (0)
error handling	How well the system identifies and addresses possible errors in the problem	– Handled all errors (4) – Handled most errors (3) – Handled some errors (2) – Handled very few errors (1) – Ignored all errors (0)
use of methods	The use of relevant techniques and concepts to address and solve the math problem.	– Excellent use (2) – Adequate use (1) – Poor use (0)
level appropriateness	How well-suited the solution provided by the system is for the given problem's level	– Highly appropriate (4) – Appropriate (3) – Moderately appropriate (2) – Slightly appropriate (1) – Not appropriate (0)
solution depth	The depth of the solution provided in terms of showing all steps and important calculations	– Extremely detailed (3) – Detailed (2) – Moderate (1) – Superficial (0)
terminology	Correct and consistent use of mathematical terminology in the explanations	– Appropriate (2) – Mostly appropriate (1) – Inappropriate (0)
reliability	The dependability of the procedure/algorithm used in providing the solution	– Distrusted (2) – Mostly Trusted (1) – Trusted (0)
calculation error	Presence of any computational or mathematical mistakes in the solution	– No errors (2) – Minor errors (1) – Major errors (0)
creativity	Novel approach or method used in providing the solution	– exceptionally novel (2) – moderately novel (1) – standard (0)
relevance	The solution should focus on solving the given problem and avoid irrelevant information or steps.	– Highly relevant (2) – Moderately Relevant (1) – Irrelevant (0)
simplification	The degree to which the solution simplifies the problem while maintaining accuracy	– Completely (3) – Mostly (2) – Partially (1) – Not at all (0)
handling constraints	The accuracy of the solution in addressing given constraints	– Fully respected (2) – Partially respected (1) – Disregarded (0)
problem type	The type of the math problem	– Excellent (4) –Good (3) – Average (2) –Poor (1) – Terrible (0)
adaptability	Adaptability refers to the ability of the solution provided to be modified and adjusted to alternative or related problems.	– Other (11) – Logic (10) – Topology (9) – Differential Equations (8) – Linear Algebra (7) – Number Theory (6) – Combinatorics (5) –Statistics (4) – Calculus (3) –Trigonometry (2) – Geometry (1) –Algebra (0)
problem level	The difficulty level of the math problem	– Level 5 (4) –Level 4 (3) – Level 3 (2) –Level 2 (1) – Level 1 (0)
solution approach	Appropriateness of the solution approach used	– Appropriate (2) – Questionable (1) – Inappropriate (0)
correct reasoning	The extent to which the systems response demonstrates correct mathematical reasoning.	– 100% (4) – 75% (3) – 50% (2) – 25% (1) – 0% (0)

Table 5: Comprehensive Verification Criteria for ALF-World Housholding Tasks.

Criteria	Description	Accepted Values
task completion	Degree to which the task is completed successfully	– 100% (4) – 75% (3) – 50% (2) – 25% (1) – 0% (0)
action validity	Actions must be chosen from the given available actions, with illegitimate actions taken into account	– all_legal (3) – one_illegal (2) – two_illegal (1) – three_or_more_illegal (0)
thought process	The quality of the thought process and planning throughout the task	– excellent (3) – good (2) – fair (1) – poor(0)
systematic search	How systematically the player searched for items and target locations	– excellent (3) – good (2) – moderate (1) – poor (0)
interaction flow	The smoothness and continuity of interactions with the environment	– smooth (2) – some_disruptions (1) – frequent_disruptions (0)
task time	The time taken to accomplish the task	– very_fast (3) – fast (2) – average (1) – slow (0)
planning strategy	Quality of the devised plan for completing the task	– excellent (3) – good (2) – fair (1) – poor (0)
action efficiency	Efficiency of the chosen actions in solving the task	– very_high (3) – high (2) – moderate (1) – low (0)
response format	Adherence to the required response format	– correct (2) – partially correct (1) – incorrect (0)
adaptability to feedback	Ability to adapt and modify the plan based on the environment's feedback	– very_high (3) – high (2) – moderate (1) – low (0)
termination judgement	Proper judgment of when to reply with 'TERMINATE'	– correct (2) – partially correct (1) – incorrect (0)
efficiency	Assesses the number of steps taken in comparison to the minimum possible steps required to complete the task	– optimal (3) – near_optimal (2) – average (1) – below_average (0)
problem solving	The ability to quickly identify and adapt to changes in the environment during task execution	– fast_adaptation (3) – moderate_adaptation (2) – slow_adaptation (1) – no_adaptation (0)
target handling	How well the player followed instructions for handling the target household	– excellent (3) – good (2) – moderate (1) – poor (0)
environment understanding	The ability to understand the provided environment description and identify relevant objects	– excellent (3) – good (2) – fair (1) – poor (0)
compliance with instructions	Adherence to specific rules and instructions such as reply formatting and termination conditions	– compliant (2) – partially compliant (1) – non-compliant (0)
legal actions	Selecting actions from the given available actions and avoiding illegal actions	– excellent (4) – good (3) – average (2) – below_average (1) – poor (0)
target acquisition	Acquiring the target household object immediately after finding it	– excellent (3) – good (2) – fair (1) – poor (0)
format adherence	The extent to which the output format is strictly followed	– Correct format (2) – Minor format issues (1) – Incorrect format (0)
problem understanding	Understanding of the given task and relevance of the environment	– 3 - Fully understood (3) – 2 - Adequately understood (2) – 1 - Partially understood (1) – 0 - Not understood (0)
action selection	Choosing the appropriate sequence and type of actions	– 3 - Optimal selection (3) – 2 - Good selection (2) – 1 - Somewhat acceptable selection (1) – 0 - Poor selection (0)