AUTOLIBRA: AGENT METRIC INDUCTION FROM OPEN-ENDED HUMAN FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Agents are predominantly evaluated and optimized via task success metrics, which are coarse, rely on manual design from experts, and fail to reward intermediate emergent behaviors. We propose *AutoLibra*, a framework for agent evaluation, that transforms open-ended human feedback *e.g.* "*If you find that the button is disabled, don't click it again*", or "*This agent has too much autonomy to decide what to do on its own*" into metrics for evaluating fine-grained behaviors in agent trajectories. AutoLibra accomplishes this by grounding feedback to an agent's behavior, clustering similar positive and negative behaviors, and creating concrete metrics with clear definitions and concrete examples, which can be used for prompting LLM-as-a-Judge as evaluators. We further propose two *meta-metrics* to evaluate the alignment of a set of (induced) metrics with open feedback: "coverage" and "redundancy". Through optimizing these meta-metrics, we experimentally demonstrate AutoLibra's ability to induce more concrete **agent evaluation** metrics than the ones proposed in previous agent evaluation benchmarks and discover new metrics to analyze agents. We also present two applications of AutoLibra in **agent improvement**: First, we show that AutoLibra serve human prompt engineers for diagonalize agent failures and improve prompts iterative. Moreover, we find that AutoLibra can induce metrics for automatic optimization for agents, which makes agents improve through self-regulation. Our results suggest that AutoLibra is a powerful task-agnostic tool for evaluating and improving language agents.

1 INTRODUCTION

Humans readily acquire skills from open-ended instructions and feedback from others (Tomasello et al., 1993). These instructions and feedback are internalized for self-regulated learning (Pintrich & Zusho, 2002; Nicol & Macfarlane-Dick, 2006), providing internal signals for continuous improvement. Drawing inspiration from this process, we investigate how well AI agents can benefit from open-ended human feedback through induction of generalizable metrics.

In this paper, we introduce AutoLibra, a metric induction method, as a novel agent evaluation framework that mitigates the limitations of current evaluation paradigms. AutoLibra is an evaluation tool that induces interpretable metrics for AI agents from open-ended human feedback, which can be collected from end users of AI agents or experts. This offers two advantages: (1) It is much easier to provide concrete feedback for trajectories than creating metrics, and (2) AutoLibra allows us to evaluate agents from the perspective of the users. AutoLibra-induced metrics provide concrete definitions of behaviors that the model-based evaluation method should look for, which could be used to understand agent behavior, as well as optimization targets to improve agents.

Inspired by the code-theme steps of thematic analysis conducted by experts in social sciences (Braun & Clarke, 2006), we design the AutoLibra induction process (§2.2) as two steps: (1) *feedback grounding*: where we ground every aspect of human feedback on some behavior in the entire agent trajectory, and (2) *behavior clustering*: where we cluster the aspects into multiple clusters of similar behaviors to summarize into metrics. As illustrated in Fig. 1, the user gives a web agent feedback "the agent did not choose iPhone 14/15" which is grounded to the agent's behavior, choosing "iPhone 16 Pro" from the drop-down menu. Similar behaviors are clustered into a common cluster, summarized as *Element Interaction Accuracy*.

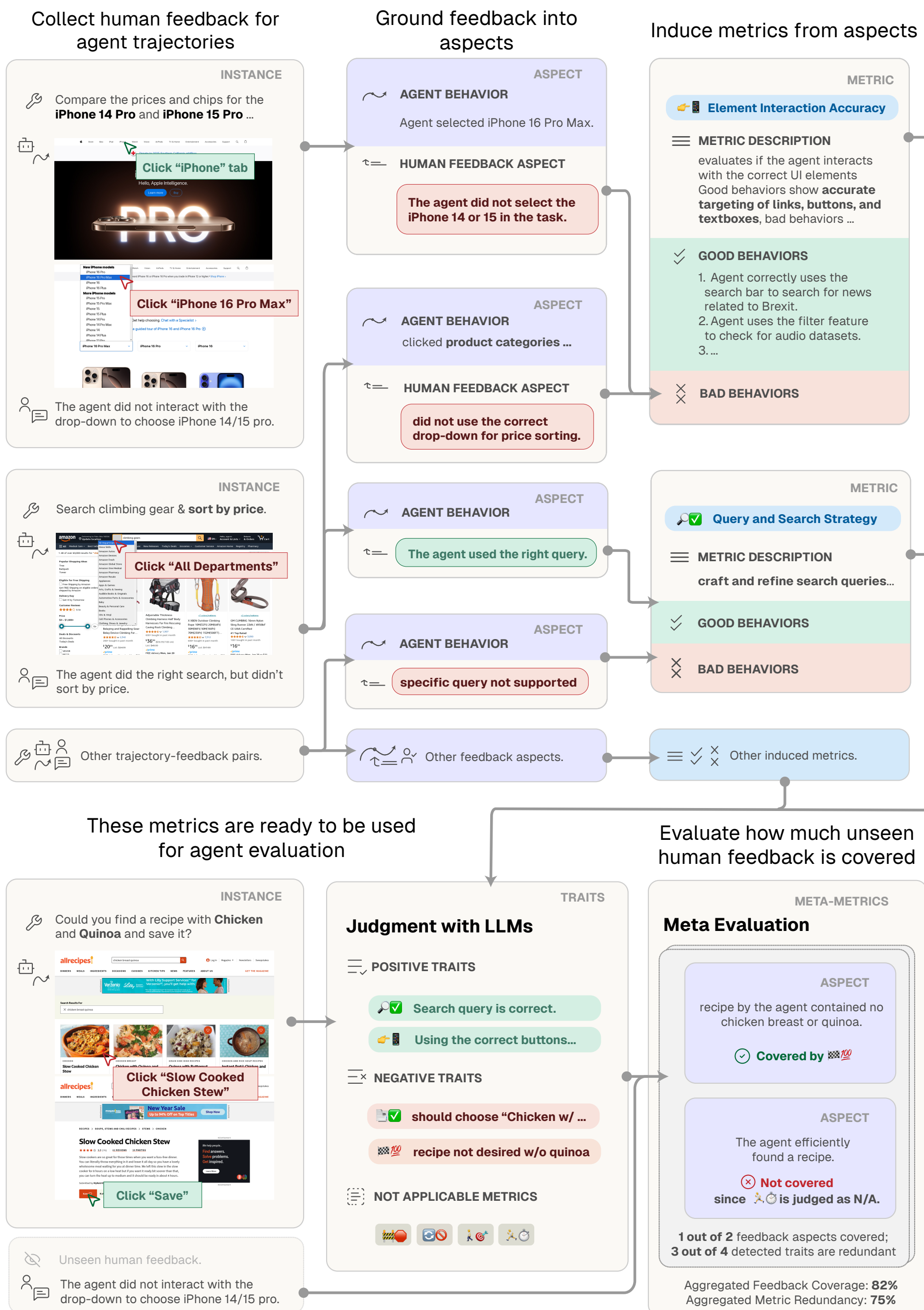

Figure 1: AutoLibra ⚖ induces agent evaluation metrics from human feedback, and uses these metrics to evaluate agents, which can be meta-evaluated via evaluating the coverage on unseen human feedback. Here we show real examples of agent trajectories, human feedback, aspects, induced metrics, evaluation results on WebVoyager (He et al., 2024).

The AutoLibra evaluation process is designed to provide a closed-loop feedback signal for the induction process. The agent trajectories used in the induction process are scored by LLM-as-a-Judge (Zheng et al., 2023) on the induced metrics. The evaluation process (§2.3) then tries to match the feedback aspects, *e.g.* "recipe does not contain quinoa", with the traits, *e.g.* `task-requirement-achievement`. In this way, we can meta-evaluate the quality of the metrics: (i) *coverage* (what proportion of feedback aspects can be matched with an agent trait), and (ii) *redundancy* of the metrics (what proportion of the detected traits are not mentioned by humans).

These two metrics provide an overall statistical picture of the quality of the induced metrics. Based on these two metrics, we can search for the set of metrics with the lowest redundancy within those with the highest coverage. As shown in §3.1, we find that as the number of metrics increases, the redundancy increases, and the coverage ultimately converges to the maximum coverage. With AutoLibra, our aim is to answer the following research questions:

RQ1: How well do AutoLibra's step-wise results align with human judgment?
RQ2: Does AutoLibra provide insights into agent behavior beyond expert-designed metrics?
RQ3: Can AutoLibra provide optimization signals for improving agents' performance?

Experiments within multiple agent domains, including collaborative agents (Shao et al., 2024), social agents (Zhou et al., 2024b), web agents (Zhou et al., 2024a; He et al., 2024), and text game agents (Paglieri et al., 2024; Cloos et al., 2024), demonstrate that AutoLibra is able to induce fine-grained and interpretable metrics with high coverage and low redundancy in unseen human feedback with 80 trajectories annotated with one feedback for each trajectory per dataset. These metrics are more concrete, and some of them were even overlooked in expert designed metrics or error analysis (§4). AutoLibra can iteratively discover new, emergent metrics (§3.2) throughout the agent optimization process, and provide optimization signals helps improve the performance of frontier LLM in a challenging 2D text game by over 20% (§5) in 3 stages with only 18 trajectory annotated per stage.

2 AutoLibra ⚖️

To address the limitations of existing evaluation paradigms, AutoLibra ⚖️ is designed to meet the following desiderata: (1) *induced from agent behavior*: This ensures that metrics are grounded in agent trajectories rather than predefined by human experts, (2) *self-validating*: Allows choosing minimal set of metrics that cover unseen human feedback with sufficient abstraction to be useful across different tasks, and (3) *generalizable*: Applicable to various agent environments, independent of domain-specific design. Based on feedback data collected from humans (§2.1), AutoLibra achieves these desiderata through a closed-loop pipeline consisting of two processes: **Induction Process** that converts agent behaviors and corresponding feedback into metrics, (§2.2) and **Evaluation Process** that predicts ratings and quality of new agent behaviors on the induced metrics (§2.3).

2.1 Collecting human feedback

In this paper, we use human feedback from two groups: (1) End-users – for agents that interact directly with humans, we use the feedback from the users who interact and converse with the agents. CoGym (Shao et al., 2024) is the environment that belongs to this category, and we use the user comments collected in their study, resulting in 197 trajectories with feedback. (2) Experts – for agents that do not directly interact with humans, we use the feedback from human annotators (five authors in this paper) who observe agent trajectories. All other environments belong to this category, these being Sotopia (Zhou et al., 2024b), WebArena (Zhou et al., 2024a), WebVoyager (He et al., 2024), Baba-is-ai (Cloos et al., 2024), and MiniHack (Samvelyan et al., 2021). For each trajectory, we collect only one element of feedback based on the complete agent trajectories.[1]

Annotators are instructed to explicitly indicate the aspects of agent behavior that they classify as good or bad, and to avoid general comments such as *"The agent is good at solving the task"*. The annotators can also choose from a terminal or a web interface; in both cases the annotator is provided with the agent's task and then view the agent's observation and actions step by step, in text form.[2] For multi-agent tasks, we annotate each agent's trajectory in a given interaction separately. For Sotopia (Zhou et al., 2024b), WebArena (Zhou et al., 2024a), and WebVoyager (He et al., 2024), we annotate 100 trajectories of agents based on GPT-4 (Achiam et al., 2023) with feedback for each dataset. For experiments in §5 we annotate 18 trajectories for each dataset in each iteration. The annotation process is fast: Human annotators spend less than 5 minutes to provide feedback for each trajectory; §4, we randomly hold out 20% of the trajectories for validation.

[1]While in theory we can leverage feedback on specific steps to achieve better feedback grounding and multiple feedback for single trajectory, we leave it as future work.

[2]While viewing screenshots is standard for web navigation tasks, we keep the observation format consistent across agents and humans to encourage more grounded feedback.

2.2 INDUCTION PROCESS

Feedback Grounding The feedback of human annotators can contain multiple aspects; e.g. *"AI agent was pretty good at giving me a consistent itinerary and vacation plan, although it froze on the last couple of minutes."*, collected from human annotators in CoGym (Shao et al., 2024), contains a positive aspect about the agent's ability to generate a consistent itinerary, and a negative aspect about the agent freezing at the end. Here we define an *aspect* as a triple (`behavior`, `feedback`, `sign`). In the positive aspect of the previous example, the `behavior` is the agent's actions to create a 20-day itinerary for the Maldives, the `feedback` is that the created itinerary is consistent and the `sign` is positive. This grounding procedure is similar to the coding procedure in thematic analysis.

We feed the trajectory and the feedback into the LLM (we use GPT-4o (OpenAI et al., 2024) as it yields good results in our pilot experiments) and prompt the LLM with the following instructions: (1) break down the feedback into bullet points; (2) for each bullet point, find the corresponding part of the trajectory to which the feedback refers. Finally, we use constrained decoding to force GPT-4o to output the aspects in the previous format. In our experiments, we find that on most datasets, for each trajectory, the LLM can generate one to five aspects, with a mean of one to two aspects.

Behavior Clustering The second step of the extraction process is to group the aspects into N metrics. To illustrate this step, we consider another example in the same dataset *"The AI responds quickly to write and run the Python script"* where the `behavior` is the agent's action to quickly write and run a Python script, the `feedback` is that the agent responds quickly, and the `sign` is positive. Although this aspect is a positive aspect, it reflects the same dimension of the agent's behavior as the previous negative aspect, with an opposite value. Each *metric* is a cluster of aspects, with a definition summarizing the criteria of positive behaviors, a list of positive behavior examples, and a list of negative behavior examples. This clustering procedure is similar to the theme induction step in thematic analysis.

However, clustering similar agent behaviors together is challenging for statistical clustering methods.[3] Inspired by LLM-based semantic clustering and concept induction methods Viswanathan et al. (2024); Lam et al. (2024), we prompt an LLM (o3-mini high[4], as it produces the most accurate coverage and redundancy scores as evaluated later) to cluster the aspects into metrics. As illustrated in Fig. 6, we gather all the aspects of M trajectories and cluster into N metrics, where N is a parameter set through the optimization process (§3.1). We provide the LLM with the following instructions: *The granularity of the grouping should be minimal; only very similar behaviors are grouped together; but don't limit to one particular website or one particular character*, which empirically makes the metrics more concrete but still applicable across different tasks.

2.3 EVALUATION PROCESS

Evaluating agents with induced metrics LLM-as-a-Judge (Zheng et al., 2023), or more broadly, model-based evaluation (Zhang et al., 2019; Celikyilmaz et al., 2021) is a method to use machine learning models to evaluate the output of other machine learning models. The success of LLM-as-a-Judge depends on the gap between the difficulty of evaluation or verification and that of generation and action. In agentic tasks, this gap is often large, as the policy model must perform multiple steps in decision-making, while the evaluation model must only classify the trajectories, which make LLM-as-a-Judge widely used (Zhou et al., 2024a; He et al., 2024; Zhou et al., 2024b). In AutoLibra, we employ LLM-as-a-Judge to evaluate the agent trajectories configured with the induced metrics. However, LLM-as-a-Judge can be replaced by any other evaluation methods implementing the induced metrics; *e.g.* an `interact-valid-element` metric could be evaluated by a rule-based evaluator that checks if the agent interacts with valid elements on the webpage. Wenote that AutoLibra could be used with other evaluation methods, such as programmatic evaluation (Ma et al., 2024); we leave generating programs for the induced metrics for future work.

As illustrated in Fig. 7, taking the induced metrics as input, an LLM (we use o3-mini medium, as it provides similar results in this step to o3-mini high) is prompted to rate the agent trajectories to

[3]In preliminary experiments, we tried to use K-means clustering on the aspect vectors generated by `text-embedding-3-large`, but the clusters are mostly based on tasks and not on the behaviors.

[4]https://openai.com/index/openai-o3-mini/

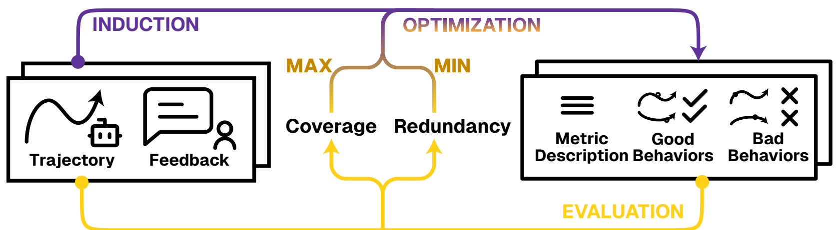

Figure 2: Metric optimization: optimizing the induction process through maximizing the coverage while minimizing redundancy of the metrics, calculated via the evaluation process.

{+ 1, -1, N/A} for each metric. For an agent trajectory, the metrics labeled +1 are the positive *traits*, and the ones labeled -1 are the negative *traits*. When we calculate the scores of the metrics, we use the ratio of agent trajectories rated as positive to the ones that are rated as positive or negative, ignoring those rated as N/A, since not all metrics are applicable to all trajectories (some metrics like `valid-search-terms` are only applicable when the task involves searching).

Meta evaluation The final loop component is the meta-evaluation, i.e. evaluating the evaluation metrics induced by AutoLibra. This step matches the traits detected by the LLM-as-a-Judge with aspects grounded from the human feedback. The goal is to verify whether (1) the induced metrics cover the behaviors the human annotators care about, and (2) LLM-as-a-Judge can produce accurate evaluation results based on the induced metrics. In the previous example, if the `respond-promptly` is extracted as a metric, and the LLM-as-a-Judge has the same opinion as the human annotators, then this aspect is considered as successfully covered. If either a similar metric was not extracted, or the LLM-as-a-Judge assigns a different score, then this aspect is considered as not covered.

As illustrated in Fig. 8, we perform meta-evaluation for each trajectory-feedback pair by classifying the aspects into positive and negative aspects, classifying traits into positive and negative traits based on rating, then matching the positive aspects with positive traits and the negative aspects with negative traits. We prompt an LLM (we use GPT-4o (OpenAI et al., 2024)) with a list of aspects and another list of traits and ask the LLM to find the best matching trait for each aspect or decide that there is no matching trait. The *coverage* of the whole dataset is calculated as the proportion of aspects of all instances that have a matching trait, and the *redundancy* is calculated as the proportion of traits of all instances that have not been matched with any aspect.

3 Optimizing and Validating AutoLibra ⚖️

AutoLibra is designed to be self-validating through the evaluation process, which allows us to search the optimal set of metrics that cover the human opinion the best (§3.1). This optimization process can also be applied iteratively throughout the agent improvement process. As the agent is optimized, new metrics can be added to existing metrics (§3.2), which is similar to how unit tests are kept throughout software development to prevent new features from interfere with existing features. In the last part of this section, we study the alignment between each step of AutoLibra and human judgment.

3.1 Metric Optimization

As illustrated in Fig. 2, we optimize the metric induction process to maximize **coverage** and minimize **redundancy**. Among the two, we prioritize coverage of the metrics to provide a comprehensive evaluation of the agent behavior, while minimizing overlap within the metrics to avoid redundancy, thus maximizing the utility of induced metrics. To optimize for this objective, we generate 20 different sets of metrics, with metric count N ranging from 4 to 13, and calculate the coverage and redundancy of the metrics in human feedback. We then select metrics with a coverage of at least the highest coverage minus 1%, and the lowest redundancy. This is performed iteratively, by resetting the range of N to the number of metrics selected previously ± 2, repeating until the coverage and redundancy of the selected metrics converge, normally within 3 iterations. While this optimization process is simple, experiments with various other optimization strategies, including genetic algorithms and iterative clustering saw none of them yield better results than the simple strategy. Fig. 3 shows the highest coverages of the metrics of size N, which converge around $N = 6$ to 10 depending on the datasets. The best coverage on Sotopia (Zhou et al., 2024b) is the lowest among all four datasets, 60%, likely due to the diversity of the tasks in the dataset, while coverage on WebArena (Zhou et al., 2024a) and WebVoyager (He et al., 2024) are the highest, 88%. We also find that the coverage of the held-out trajectories is only slightly worse ($< 5\%$) than the trajectories we use to induce the metrics, which is expected since we use the exact examples extracted from the latter. Lastly, we show that the good and bad behaviors are crucial in the metrics, dropping which resulting in up to 30% coverage decrease on CoGym.

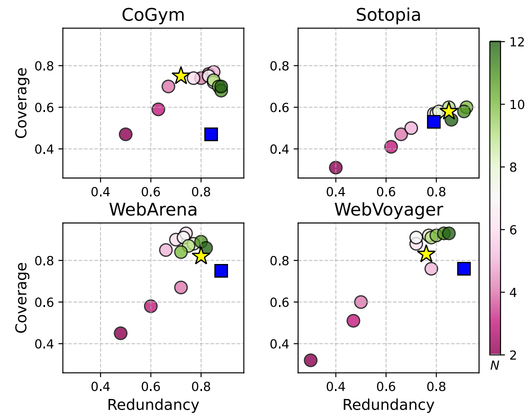

Figure 3: Coverage and redundancy of AutoLibra metrics on four agentic datasets. Circles indicate coverage and redundancy for different induced metrics; stars indicate the the best metrics' coverage and redundancy on held-out human feedback; squares show an ablation test, indicating the effect when good and bad behavior examples are removed from metrics, demonstrating the criticality of concrete behavior examples

3.2 Iterative Metric Induction

When applying AutoLibra to agent optimization, we can iteratively induce new metrics, as agents develop new failure modes or new behaviors as they improve, which is useful for tracking agents' progress across different iterations.[5] To do this, we modify the behavior clustering step, by providing the LLM with the existing metrics and their definitions, and ask the LLM not to change the definitions of the existing metrics, to only add new behaviors to the existing metrics, and add new metrics if necessary. We apply the same optimization strategy as in the metric optimization step ensure the newly induced metrics cover emerging behaviors and do not overlap with existing metrics.

Table 1: The ratio of instances marked as fully correct in human validation. For each step and each task, we randomly sample 40 instances to reach a relatively small confidence interval of 0.04 and ask human annotators to label them as completely correct or not. Although the agreement scores vary across tasks and steps, the average agreement for each step and dataset is above 0.85 significantly.

Steps	CoGym	Sotopia	WebArena	WebVoyager	Baba-is-AI	Average
Grounding	0.95	0.95	0.98	0.93	0.93	0.95 (±0.03)
LLM-as-a-Judge	0.90	0.85	0.95	1.00	0.90	0.92 (±0.04)
Meta-Evaluation	0.98	0.90	0.85	0.83	0.95	0.90 (±0.04)

3.3 How aligned are the steps in AutoLibra with human judgment?

Since AutoLibra uses LLMs in each step, we first ask whether LLM outputs are reliable or aligned with human judgment. To measure the alignment of AutoLibra metric induction with human judgment, we validate the feedback grounding, agent evaluation, and meta evaluation steps by having human experts manually review each step (with exception of the behavior clustering step, as it is prohibitively time-intensive for human annotators to process and cluster more than 400 aspects), scoring (1/0) based on whether they agree with the outcomes of each iteration. The coverage and redundancy scores, in combination with the validation results of the other steps in the loop, thus serve as an indirect validation for the behavior clustering step. Table 1 shows the agreement rate of human annotators in AutoLibra steps. It should be noted that these tasks are significantly different; e.g., grounding for WebVoyager (He et al., 2024) is challenging due to the length and wide action space of the trajectory, and LLM-as-a-Judge for Sotopia (Zhou et al., 2024b) is difficult due to the complexity of the evaluation of social interactions. Our results show that the majority (significantly over 85%) of results in AutoLibra are reliable according to human validation.

[5]Alternatively, a new set of metrics can be induced from scratch for each iteration - in practice, we do not find that this results in any coverage loss, but we choose the former method for consistency

4 AutoLibra as a lens : agent evaluation with AutoLibra

In this section, we use AutoLibra as a lens to provide grounded, behavior-salient insights into agent trajectories. In three data sets, CoGym (Shao et al., 2024), Sotopia (Zhou et al., 2024b), and WebVoyager (He et al., 2024), we compare induced metrics with heuristically proposed evaluation dimensions and failure modes summarized by the authors. We find that AutoLibra can discover more concrete metrics than heuristically defined categories, and novel metrics that are overlooked by experts. Tab. 2 summarizes the comparison between AutoLibra-induced metrics and evaluation criteria across the three aforementioned datasets. Check out detailed analysis in App. §B.

For CoGym (Shao et al., 2024), AutoLibra induces 9 metrics from end user feedback that correspond to the five failure categories proposed by authors, with failure rates matching manually labeled categories and providing automated measurement of agent failures. For Sotopia (Zhou et al., 2024b), AutoLibra recovers the exact *Goal Completion* dimension and three subdimensions of *Believability*, while discovering four additional metrics overlooked in the original design. AutoLibra minimizes redundancy by consolidating overlapping dimensions into a single *Goal Achievement and Outcome Effectiveness* metric. For WebVoyager (He et al., 2024), AutoLibra discovers concrete behavioral metrics such as *Access Barrier Handling*, *Error Recovery and Adjustment*, and *Navigation Accuracy* that provide more specific characterization than previous "navigation stuck" classifications (He et al., 2024; Zhou et al., 2024c). The framework identifies additional failure modes like *Query Strategy Efficiency* (7%) and *Final Output Quality* (18%) not captured in prior analyses.

Table 2: AutoLibra-induced metrics and expert-proposed evaluation dimensions and failure categories. Percentages in parenthesis denote failure frequency or score from AutoLibra or the original papers.

	AutoLibra-induced metrics	**Failure categories by experts**
	Matched metrics and failure categories	
CoGym (Shao et al., 2024)	*Responsiveness and Efficiency* (75%) *Communication Clarity & Notification* (8%)	*Communication* (65%)
	Instruction Adherence & Follow-Through (24%)	*Situational Awareness* (40%)
	Iterative Refinement and Adaptability (47%) *Autonomy and Proactiveness* (28%)	*Planning* (39%)
	Content Quality and Coherence (16%) *Search and Retrieval Accuracy* (13%) *Data Analysis Competence* (2%)	*Environmental Awareness* (28%)
	Interface and User Experience (23%)	*Personalization* (16%)
	Matched metrics and social dimensions	
Sotopia (Zhou et al., 2024b)	*Goal Achievement & Outcome Effectiveness* (19%)	*Goal Completion* (14%)
	Conversational Naturalness & Efficiency (5%) *Personality Consistency and Alignment* (2%) *Contextual Integration of Identity* (1%)	*Believability* (4%)
	Unmatched AutoLibra-induced metrics	
	Negotiation Tactics and Strategic Adaptability (14%), *Responsiveness and Conversational Termination* (5%), *Adaptability and Flexibility in Dialogue* (7%)	
	Unmatched Sotopia-Eval dimensions	
	Relationship, Knowledge, Secret, Financial and Material Benefits, Social Rules	
	Matched metrics and failure reasons	
WebVoyager (He et al., 2024)	*Error Recovery & Adjustment* (15%) *Step Efficiency & Action Redundancy* (13%) *Navigation Accuracy* (11%) *Access Barrier Handling* (2%)	*Navigation Stuck* (44%)
	Information & Verification Accuracy (16%)	*Hallucination* (22%)
	Result Relevance Accuracy (9%)	*Prompt Misalignment* (9%)
	Unmatched AutoLibra-induced metrics	
	Query and Search Strategy Efficiency (7%), *Final Output and Summarization Quality* (18%)	
	Unmatched WebVoyager fail reasons	
	Visual Grounding Issue (25%)	

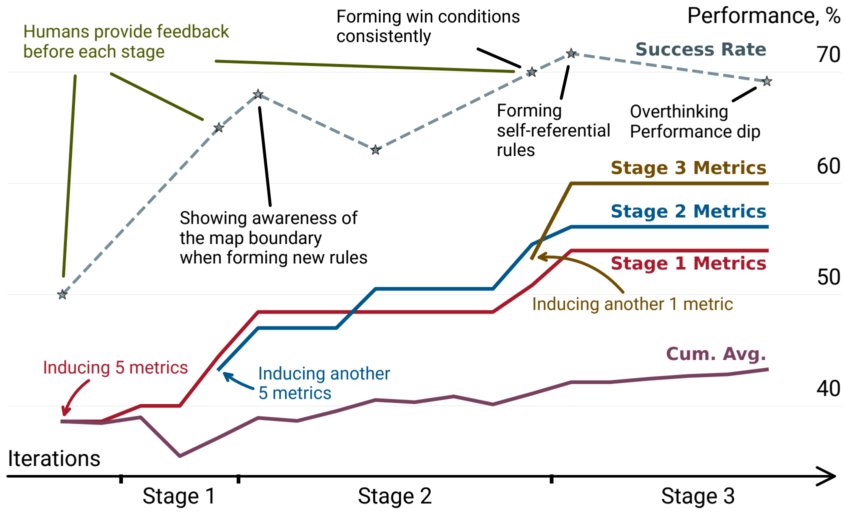

Figure 4: AutoLibra iteratively induce metrics and improves the agent prompts through optimizing for the induced metrics. Although not optimized for, the success rate of the agent continuously improve until Stage 3, when the agent begins to overthink.

5 AutoLibra as a ladder : Agent improvement with AutoLibra

As AutoLibra can automatically induce metrics from human feedback, a natural question to ask is whether it can enable self-regulated improvement in agents through iterative feedback. This can be achieved through optimizing the agent prompts towards higher scores on the metrics extracted by AutoLibra. To answer this question, we use a challenging 2D game Baba-Is-AI (Cloos et al., 2024; Paglieri et al., 2024) as a benchmark. Inspired by Baba-Is-You, this game requires not only following rules to achieve goals, but also manipulating the rules, even self-referential ones. For example, in the game illustrated in App. Fig. 9, the agent needs to change self-referential rules from *baba is you*, to *door is you* to control the green door on the other side of the wall, form a new win rule *ball is win*, and navigate to the red ball to achieve the win condition. To achieve a high score on this dataset, the agent needs not only planning, but also metacognitive skills, which is very challenging for LLM agents with frontier models as shown in the Balrog benchmark (Paglieri et al., 2024). In this experiment, we use Gemini-2.5-Flash (Team et al., 2025) for the agent, AutoLibra, and agent prompt optimization, throughout the experiment, which will be referred as the LLM in this section. Gemini-2.5-Flash is ranked as the 3rd place, with a success rate of $50.8\% \pm 4.6\%$ on the Balrog leaderboard for Baba-is-AI at the time of submission, and the state-of-the-art result is $56.7\% \pm 4.5\%$.

Fig. 4 illustrated our procedure, and summarized the results. We employ an iterative process by improving the agents in 3 stages through providing human feedback on 6 out of 40 tasks in the Baba-Is-AI. Before each stage we show human annotators 3 trajectories for the 6 tasks, gather the feedback, and apply AutoLibra iterative metric induction process (§3.2). This results in 5 metrics for Stage 1 and 2, and another 1 metric for Stage 3. Within each stage, we iteratively feed 1 LLM agent trajectory on each of these 6 tasks, together with evaluation results based on these AutoLibra-induced metrics to the LLM to improve the prompt of the LLM agent. This process results in continuous improvement not only on the running maximum metric scores, the cumulative average metrics, but also game success rate. Fig. 4 shows these statistics on the whole 40 tasks, although we only use 6 out of the 40 tasks in the whole optimization process. Upon examining the agent trajectories, we find the skills learned in the process. In the first stage, the agent learns to find rules to form based on the map boundary, which could be a result of an induced metric `map-n-constraint-recognition`. Similarly, more advanced skills are learned in Stage 2 and 3, including forming win conditions and self-referential rules, probably as a result of metric `rule-manipulation-proficiency`.

Our results show that the metrics induced by AutoLibra form effective objectives for improving the agents through prompt optimization. It should note that AutoLibra is a metric induction method, which is orthogonal to learning algorithms, including prompt optimization, fine-tuning or reinforcement learning. We show that this process improves agent success rate by 20% without optimizing for success rate, and in the future, researchers can study the effect of employing other learning algorithm.

6 Related Work

AutoLibra unifies three areas of research: it draws inspiration from *thematic analysis* to create *nautral language-derived evaluation metrics* to evaluate and reward *AI agents*.

Evaluating AI agents Much of the work in AI agent evaluation focuses around benchmarks which contains both task suites and evaluation metrics. In addition to the datasets we used in this paper, SWE-Bench (Jimenez et al., 2024) uses human-written unit tests as evaluation metrics; Embodied Agent Interface (Li et al., 2024) provides fine-grained evaluation for LLM-based embodied agents; τ-Bench (Yao et al., 2024) compares database states for evaluation; concurrent work AgentRewardBench (Lù et al., 2025) builds a benchmark for reward models for web agents. Recently, there are observatory tools including Galileo (Galileo, 2025), Vertex AI Gen AI (Cloud, 2025), and Docent (Meng et al., 2025) which provide user interfaces to visualize agent failure modes. Generating intrinsic rewards have also been studied in the reinforcement learning community (Du et al., 2019; Pathak et al., 2017; Laskin et al., 2022) to encourage exploration, sub-task completion, or skill discovery. In contrast to these, AutoLibra is a pure data-driven task-agnostic method without predefined failure taxonomy for generating interpretable metrics for agents.

Learning from natural language and human feedback Researchers have been studying reinforcement learning with language feedback to provide a dense reward to agents (Goyal et al., 2019). Since LLM agents are even harder to train with sparse reward, there is substantial interest in training LLM agents from natural language feedback. Chen et al. (2024) propose an imitation learning method for learning from human feedback; Text2Reward (Xie et al., 2024) uses code generation to generate robot reward functions from open-ended human feedback; our work (Chen et al., 2025) uses feedback to the improvement agent policy with prompting and then align the unprompted agent policy with the prompted one; Shi et al. (2024) propose a new model architecture to incorporate human feedback into policy learning. On the other hand, human non-open-ended feedback is also incorporated in training agents, including rating feedback (Nguyen et al., 2017), preference feedback (Christiano et al., 2017), demonstrative feedback (Shaikh et al., 2025). Unlike these papers, AutoLibra induces metrics from feedback from all annotated instances and generates metrics that are generalizable to different tasks and useful for both evaluation and agent fine-tuning.

Thematic analysis Thematic analysis is a powerful tool for qualitative study through coding and iterative creation of themes. Gauthier & Wallace (2022) provide computational tools to aid this process; Hong et al. (2022) and Gebreegziabher et al. (2023) explore human-AI collaboration in thematic analysis; LLooM (Lam et al., 2024), an automatic method for concept induction, closly aligns with and informs our approach. This paper completes the loop of concept induction by using the meta-evaluation step to optimize the induced metrics, and apply it to agent evaluation.

7 Conclusion and Future Work

This work introduces AutoLibra, a new paradigm for agent evaluation, one of the first works to explore adaptable trajectory-derived evaluation heuristics, offering substantial advantages in agent training over traditional end-to-end evaluation. We find that this framework is generalizable to a diverse range of agent tasks, provides new insights into agent behaviors, and identifies strong optimization targets for agent improvement. There are a few directions for further extending and applying this framework. (1) **Behavior-centric evaluation** AutoLibra leads a *paradigm shift* from end-to-end agent evaluation (analogous to “integration tests” in software development) to evaluation with granular metrics that measure agents’ concrete behaviors (analogous to “unit tests”). Future work can study whether this process can be improved through better human-AI collaboration. (2) **Sub-trajectory feedback from humans** In AutoLibra, we label each trajectory with one piece of feedback, and ground it into the agents’ concrete behavior which is at the sub-trajectory level. In the future, researchers can let users directly give feedback for one or multiple steps in the trajectory, which should lead to better feedback grounding results. Similarly, user feedback can be collected during the interaction instead of after the agent has completed the tasks, which is a more user-friendly way to gather high quality feedback data. (3) **Wider exploration of agent improvement methods** In this paper, we only explored non-parametric for agent improvement to show the utility of AutoLibra. Future work can use AutoLibra to provide dense rewards for individual steps, and use reinforcement learning to train agents with these dense rewards.

ETHICS STATEMENT

This research adheres to the ICLR Code of Ethics. Within human-aided experiments, we are also limited by the diversity of human annotators. The annotation of the data in this paper, are performed through objective and blinded surveys filled out by the authors who do not know which models that they are annotating. The human feedback for CoGym (Shao et al., 2024) is published by the original authors. Since the annotations are objective surveys on the performance of the agents without any harm to the authors or personal information gathered, this is exempted from IRB review based on the policy of authors' institution.

REPRODUCIBILITY STATEMENT

To ensure reproducibility of our results, we provide comprehensive documentation of our experimental setup and methodology in the appendix of our work. All experimental details, including model configurations, prompting strategies, and evaluation metrics, are specified in the relevant sections and supplementary materials. All code and data will be available upon acceptance.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey, 2021. URL https://arxiv.org/abs/2006.14799.

Angelica Chen, Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Samuel R. Bowman, Kyunghyun Cho, and Ethan Perez. Learning from natural language feedback. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=xo3hI5MwvU.

Wentse Chen, Jiayu Chen, Fahim Tajwar, Hao Zhu, Xintong Duan, Russ Salakhutdinov, and Jeff Schneider. Fine-tuning llm agents with retrospective in-context online learning. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Nathan Cloos, Meagan Jens, Michelangelo Naim, Yen-Ling Kuo, Ignacio Cases, Andrei Barbu, and Christopher J. Cueva. Baba is ai: Break the rules to beat the benchmark, 2024. URL https://arxiv.org/abs/2407.13729.

Google Cloud. Introducing agent evaluation in vertex ai gen ai evaluation service, 2025. URL https://cloud.google.com/blog/products/ai-machine-learning/introducing-agent-evaluation-in-vertex-ai-gen-ai-evaluation-service. Accessed: 2025-04-24.

Yali Du, Lei Han, Meng Fang, Ji Liu, Tianhong Dai, and Dacheng Tao. Liir: Learning individual intrinsic reward in multi-agent reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Galileo. Introducing agentic evaluations, 2025. URL https://www.galileo.ai/blog/introducing-agentic-evaluations. Accessed: 2025-04-24.

Robert P Gauthier and James R Wallace. The computational thematic analysis toolkit. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–15, 2022.

Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L Glassman, and Toby Jia-Jun Li. Patat: Human-ai collaborative qualitative coding with explainable interactive rule synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2023.

Prasoon Goyal, Scott Niekum, and Raymond J Mooney. Using natural language for reward shaping in reinforcement learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 2385–2391, 2019.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6864–6890, 2024.

Matt-Heun Hong, Lauren A Marsh, Jessica L Feuston, Janet Ruppert, Jed R Brubaker, and Danielle Albers Szafir. Scholastic: Graphical human-ai collaboration for inductive and interpretive text analysis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–12, 2022.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024.

Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–28, 2024.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Cic: Contrastive intrinsic control for unsupervised skill discovery. *arXiv preprint arXiv:2202.00161*, 2022.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2024.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stańczak, Peter Shaw, Christopher J. Pal, and Siva Reddy. Agentrewardbench: Evaluating automatic evaluations of web agent trajectories, 2025. URL https://arxiv.org/abs/2504.08942.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Kevin Meng, Vincent Huang, Jacob Steinhardt, and Sarah Schwettmann. Introducing docent. https://transluce.org/introducing-docent, March 2025.

Khanh Nguyen, Hal Daumé III, and Jordan Boyd-Graber. Reinforcement learning for bandit neural machine translation with simulated human feedback. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1464–1474, 2017.

David J Nicol and Debra Macfarlane-Dick. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2):199–218, April 2006.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, et al. Balrog: Benchmarking agentic llm and vlm reasoning on games. *arXiv preprint arXiv:2411.13543*, 2024.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

Paul R Pintrich and Akene Zusho. The development of academic self-regulation: The role of cognitive and motivational factors. In *Development of achievement motivation*, pp. 249–284. Elsevier, 2002.

Mikayel Samvelyan, Robert Kirk, Vitaly Kurin, Jack Parker-Holder, Minqi Jiang, Eric Hambro, Fabio Petroni, Heinrich Küttler, Edward Grefenstette, and Tim Rocktäschel. Minihack the planet: A sandbox for open-ended reinforcement learning research, 2021. URL https://arxiv.org/abs/2109.13202.

Omar Shaikh, Michelle S Lam, Joey Hejna, Yijia Shao, Hyundong Justin Cho, Michael S Bernstein, and Diyi Yang. Aligning language models with demonstrated feedback. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. *arXiv preprint arXiv:2412.15701*, 2024.

Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z. Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. Yell at your robot: Improving on-the-fly from language corrections. *arXiv preprint arXiv: 2403.12910*, 2024.

Mirac Suzgun and Adam Tauman Kalai. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv*, 2024. doi: 10.48550/ARXIV.2401.12954. URL https://arxiv.org/abs/2401.12954.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz

Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer

Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Michael Tomasello, Ann Cale Kruger, and Hilary Horn Ratner. Cultural learning. *Behavioral and brain sciences*, 16(3):495–511, 1993.

Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. Large language models enable few-shot clustering. *Transactions of the Association for Computational Linguistics*, 12:321–333, 2024.

Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Reward shaping with language models for reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024a.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024b.

Yifei Zhou, Qianlan Yang, Kaixiang Lin, Min Bai, Xiong Zhou, Yu-Xiong Wang, Sergey Levine, and Erran Li. Proposer-agent-evaluator(pae): Autonomous skill discovery for foundation model internet agents, 2024c. URL https://arxiv.org/abs/2412.13194.