CS329T: Agentic Research Project Planner

Shounak Ray
Stanford University
shounak@stanford.edu

Harrison Delecki
Stanford University
hdelecki@stanford.edu

1 Introduction

The body of scientific research grows extremely quickly. It is often difficult for individual researchers or small teams to stay informed amidst an ever-growing body of information. Additionally, new researchers like grad-students may require significant time to gain a level of epertise or mastery over a research area. These difficulties may negatively impact researchers' ability to compose salient, novel research directions. This difficulty may be partially addressed by research support tools that aim to aid researchers in exploring literature, identifying new research directions, and brainstorming experiments.

Large language models (LLMs) have shown promise in a variety of natural language processing tasks. They have been used to generate text, answer questions, and even write code. They have also been used to generate summaries of scientific papers. In this work, we explore the idea that LLMs may be able to generate research directions and experimental plans by combining information retrieved from scientific papers with some internal knowledge of the scientific method.

Evaluating the trustworthiness of LLMs in this context is challenging for several reasons. First, the quality of a research direction or experimental plan is subjective. In this work, we choose to evaluate the *creativity*, *completeness*, and *specificity* of the generated research directions and experimental plans. These metrics are based on the intuition that a good research direction or experimental plan should be novel, comprehensive, and actionable. Additionally, we evaluate the RAG trio of metrics to evaluate the quality of the generated text in relation to retrieved scientific papers.

In the remainder of this report is organized as follows:

- Methods. Here we provide a high-level overview of the actor and critic systems.

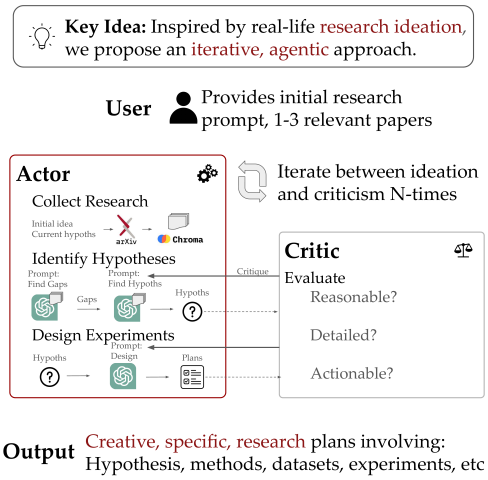

Figure 1: Illustration of our approach. The actors are responsible for collecting research, identifying hypothesis, and based on them, designing experiments. The arrows between the critics and the actors represent fine-grained critiques of the individual actors.

- Metrics and Baselines. Here we provide a description of the RAG evaluation metrics along with separate specific metrics.
- Experiments. Outlines different domains we tested the system on.
- Results and Discussion. Brief overview and main quantiatative takeaways.
- Limitations. As the name suggests.

2 Methods

In this section, we describe our approach for generating research directions and experimental plans given a user input of a few initial research papers. Inspired by the iterative nature of brainstorming, we use an actor-critic model to generate research directions and experimental plans. The

actor generates research directions and experimental plans, while the critic evaluates the quality of the generated text and provides feedback to the actor. All LLM aagents use OpenAI's GPT-4o model through API access.

2.1 Actor

The actor is responsible for generating research directions and experimental plans. The actor takes as input a few initial research papers. First, the actor LLM agent uses these papers in a RAG architechture to identify *research gaps* and *hypotheses*. From these hypotheses, we use a another LLM call without RAG to identify *research plans* designed to test each hypothesis. The research plan may consist of methods, baselines, experiments, metrics, datasets, and links to open-source tools. At each iteration, the actor expands the body of available literature by retrieving additional papers from the internet using keywords from previous hypotheses.

2.2 Critic

The critic provides fine-grained feedback to the individual actor components. Specifically, one can instrument any arbitrary function (where this function represents an individual actor agent) with "@critic.overwatch(promptfile=...)". 'promptfile' represents the name of the LLM prompt, if any, that this agent depends on to produce it's output – for instance, the 'GapFinder' agent depends on an initial prompt to help it identify research gaps from a set of documents. This will automatically register the agent into the overarching critic system and when "critic.chastise()" is run at the end of the agentic loop, the initial prompts registered into the system are updated autonomously. That is, next time the agentic loop is executed sequentially, the individual agents will operate on updated prompts to produce intermediate outputs.

3 Metrics and Baselines

In this section, we describe metrics we use to evaluate the proposed approach. We evaluate evaluate the generated research hypotheses using RAG metrics of answer relevance, context relevance, and groundedness. Answer relevance measures how well the generated text answers the user query. Context relevance measures how well the generated plan is related to the input and collected research papers. Groundedness measures how well the generated plan is grounded in the collected research. We use TruLens instrumentations to evaluate these metrics.

We also evaluate the generated research plans using our custom evaluation of creativity, specificity, and completeness. Creativity measures how novel the generated text is. Specificity measures how detailed the generated text is. Completeness measures how comprehensive the generated text is. We use LLMs-as-judges to evaluate these metrics. Prompts for these metrics are shown in appendix A.

We evaluate the proposed approach against two baselines to understand the impact of the actor-critic architecture on the quality of the generated research plans. In the simplest baseline, we simply ask ChatGPT to generate research plans without any additional information. Note that the RAG metrics are not applicable for this baseline. In the second baseline, we use an actor-only model that generates research plans without any feedback from the critic.

4 Experiments

We evaluate the system on three research areas, with three papers per area. We chose areas that we are familiar with to facilitate development. The three are black-box safety validation [1], offline reinforcement learning [2], and normalizing flows [3]. For each experiment, we evaluate metrics over 10 iterations, and perform 15 trials.

5 Results and Discussion

The complete results are presented in table 1. Overall, there appears to be a large variance in RAG metrics. Answer relevance and groundedness are relatively high, while groundedness is low. This may be because in order for the system to produce creative ideas, the answers should somehow not be entirely grounded in context, but instead draw on them and its own internal knowledge. The only RAG metric that appears to benefit from the agentic workflow is context relevance, indicating that the critic's feedback may effectively prompt the actor to pay more attention to the retrieved context. The mean answer relevance increased slightly, but the variance is quite large. Interestingly, the mean groundedness score slightly decreased when using actor-critic iterations although the variance is still quite large.

Results for the metrics over iterations are shown

Method	Answer Relevance	Context Relevance	Groundedness	Creativity	Specificity	Completeness
ChatGPT	—	—	—	0.80 ± 0.37	0.34 ± 0.26	0.57 ± 0.54
Actor Only	0.70 ± 0.40	0.92 ± 0.07	0.43 ± 0.31	0.6	0.65 ± 0.09	0.68 ± 0.10
Actor-Critic (10)	0.73 ± 0.41	0.98 ± 0.04	0.40 ± 0.36	0.6	0.76 ± 0.14	0.76 ± 0.14

Table 1: While there are marginal increases in each of the metrics above, except groundedness, moving from the baseline to the more complex implementation, the standard deviations are each quite high. While this demonstrates potential, the *quantitative* results do not allow us conclude conclusively that an agentic actor-critic system is superior than the ChatGPT baseline.

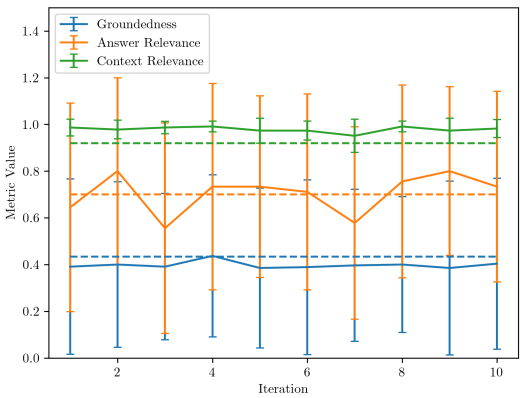

Figure 2: Demonstrates the relationship between the rag trio metrics and iterations in the agentic loop. Groundedness is unsurprisingly quite low since novelty in experimental procedures may often not be explicitly grounded in existing literature.

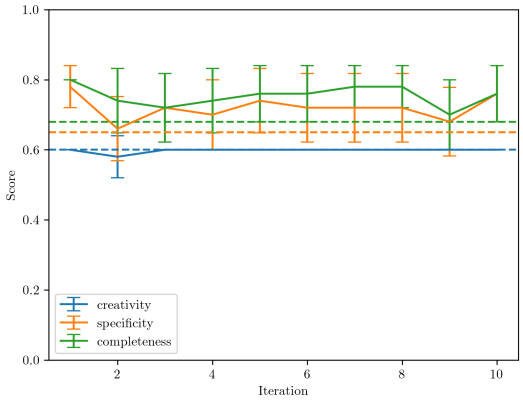

Figure 3: Demonstrates the relationship between custom metrics and number of agentic iterations. Like Figure 2, there's not a clear trend here.

in fig. 2 and fig. 3. The solid line shows the actor-critic result, while the dashed lines show the actor-only baseline. There is no clear trend in the metrics over iterations, indicating that the system may not be learning effectively from the feedback provided by the critic. This may be because the feedback provided by the critic is not actionable or because the actor is not able to effectively incorporate the feedback into its generation process.

An example experimental plan is shown in appendix B. The plan is quite detailed and includes proposed methodologies, data collection strategies, data preprocessing strategies, model architectures, evaluation strategies, potential open-source resources, evaluation metrics, baseline models, expected results, potential challenges, and a conclusion. Interestingly, this plan is quite similar to a recent popular paper on normalizing flows, indicating that the system may be able to generate high-quality research plans. However, given that the paper is from late 2023, it may be in the training set of the language model.

To improve the system, we could focus on prompt engineering, particularly in the critic. Sometimes the critic would provide feedback that was not particularly useful or actionable. Alternatively, we could fine-tune the LLMs on research papers to improve the quality of the generated research directions and experimental plans.

Known Project Limitations

1. Empirically, we get the best performance when we provide the system with a survey of a particular field along with $1 - 2$ additional papers.

2. The actor is sometimes inconsistent about the number of hypotheses/plans it generated between iterations.

3. Implementation of searching/collecting papers from arxiv is very simple, and could be improved.

4. The critiques themselves often focused on very contrived aspects of the output unrelated to the quality of the subject-matter itself (i.e. unique identifiers in JSON keys instead of content of JSON body).

5. The performance of novel qualitative metrics of creativity, specificity, completeness were highly intertwined to their respective LLM prompts.

Authorship Statement

Harrison and Shounak contributed equally to the ideation of this project. Shounak provided the architecture of the codebase. Harrison implemented actor agent, while Shounak implemented the critic. Shounak implemented metrics, and Harrison ran experiments. Both members contributed equally to the final presentation and report.

References

[1] Anthony Corso et al. "A survey of algorithms for black-box safety validation of cyber-physical systems". In: *Journal of Artificial Intelligence Research* 72 (2021), pp. 377–428.

[2] Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. "A survey on offline reinforcement learning: Taxonomy, review, and open problems". In: *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[3] George Papamakarios et al. "Normalizing flows for probabilistic modeling and inference". In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64.

A LLM Metric Prompts

Creativity: json output please. You are a hard-to-please expert evaluator tasked with assessing research proposals for creativity. Creativity in this context refers to originality, novelty, and the ability to push boundaries or explore new ideas. Given the following research blurb, assign a creativity score from 1 (not creative) to 5 (exceptionally creative) and briefly justify your rating. Remember you are hard to please, so awarding scores of 4 and especially 5 should only happen when the results are particularly good and match what you're looking for. The output should be a dictionary with two keys, one for score and another for explanation. Here is the proposed experimental design:

Specificity: json output please. You are a hard-to-please expert evaluator tasked with assessing research proposals for specificity. Specificity in this context refers to the clarity, precision, and level of detail in the research focus and methodology. Given the following research blurb, assign a specificity score from 1 (not specific) to 5 (highly specific) and briefly justify your rating. Remember you are hard to please, so awarding scores of 4 and especially 5 should only happen when the results are particularly good and match what you're looking for. The output should be a dictionary with two keys, one for score and another for explanation. Here is the proposed experimental design:

Completeness: json output please. You are a hard-to-please expert evaluator tasked with assessing research proposals for completeness. Completeness in this context refers to how well the proposal covers the necessary components for understanding its scope, objectives, and feasibility. Given the following research blurb, assign a completeness score from 1 (incomplete) to 5 (comprehensive) and briefly justify your rating. Remember you are hard to please, so awarding scores of 4 and especially 5 should only happen when the results are particularly good and match what you're looking for. The output should be a dictionary with two keys, one for score and another for explanation. Here is the proposed experimental design:

B Example Experimental Plan

Below is a proposed experimental plan produced after 3 iterations based on the survey of papers on normalizing flows. The presented idea is quite similar to the recent popular paper "Improving and generalizing flow-based generative models with minibatch optimal transport" by Tong et al. However, this paper is from late 2023, so it may be in the training set of the language model.

"Experiment Design to Test the Hypothesis: Integration of Machine Learning with Optimal Transport for Generative Models": "1. Proposed Methodologies - Framework Development: Develop a generative model that integrates optimal transport (OT) principles. This could involve using Wasserstein GANs (WGANs) as a baseline and modifying the loss function to incorporate OT metrics. Algorithm Design: Create algorithms that leverage OT for tasks such as data generation,

domain adaptation, and sample quality improvement. 2. Data Collection Strategies - Datasets: Use publicly available datasets that are commonly used for generative modeling tasks. Potential datasets include: MNIST: Handwritten digits for basic testing. CIFAR-10: A more complex dataset for image generation. CelebA: A large-scale face attributes dataset for high-dimensional data. Fashion-MNIST: Similar to MNIST but with fashion items. Synthetic Data Generation: Create synthetic datasets using known distributions (e.g., Gaussian mixtures) to evaluate the model's performance in controlled settings. 3. Data Preprocessing Strategies - Normalization: Normalize pixel values for image datasets to the range [0, 1]. Dimensionality Reduction: For high-dimensional datasets, consider using techniques like PCA or t-SNE to visualize data distributions before applying OT methods. Augmentation: Apply data augmentation techniques (e.g., rotation, flipping) to increase dataset diversity. 4. Model Architectures - Baseline Models: Wasserstein GAN (WGAN): Use as a baseline for comparison. Variational Autoencoder (VAE): Another baseline for generative modeling. Proposed Model: OT-GAN: A modified GAN that incorporates optimal transport metrics in the loss function. OT-VAE: A VAE variant that uses OT for the latent space representation. 5. Evaluation Strategies - Sample Quality Assessment: Use qualitative assessments (visual inspection) and quantitative metrics (e.g., Inception Score, Fr0̈0e9chet Inception Distance) to evaluate the quality of generated samples. Computational Efficiency: Measure training time, convergence speed, and memory usage of the proposed models compared to baseline models. 6. Potential Open-Source Resources - Software Libraries: Pytorch or TensorFlow: For implementing neural networks. POT (Python Optimal Transport): A library for optimal transport computations. Scikit-learn: For data preprocessing and evaluation metrics. Datasets: Use datasets from sources like Kaggle, UCI Machine Learning Repository, or TensorFlow Datasets. 7. Evaluation Metrics - Inception Score (IS): Measures the quality and diversity of generated images. Fr0̈0e9chet Inception Distance (FID): Compares the distribution of generated images to real images. Computational Complexity: Analyze the time complexity and memory usage of the proposed models. 8. Baseline Models - WGAN: As a baseline for generative modeling. VAE: To compare the effectiveness of OT in generative tasks. Traditional GAN: For a comparison of performance without optimal transport. 9. Expected Results - Improved Sample Quality: The OT-integrated models should produce higher quality samples than the baseline models, as measured by IS and FID. Reduced Computational Complexity: The proposed models should demonstrate faster convergence and lower memory usage compared to traditional GANs and VAEs. 10. Potential Challenges - Complexity of Implementation: Integrating OT into existing generative models may require significant adjustments and optimizations. Hyperparameter Tuning: Finding the right hyperparameters for the OT-based models may be challenging and time-consuming. Scalability: Ensuring that the proposed methods scale well with larger datasets and higher dimensions could be a potential issue. Interpretability: Understanding how the integration of OT affects the generative process may require additional analysis. Conclusion This experiment aims to rigorously test the hypothesis that integrating machine learning with optimal transport can enhance generative modeling. By following the outlined methodologies, data strategies, and evaluation metrics, we can assess the effectiveness of the proposed framework and its potential contributions to the field of generative models."