Science in the Age of AI

Chenhao Tan

Chicago Human+AI Lab

University of Chicago

https://chenhaot.com

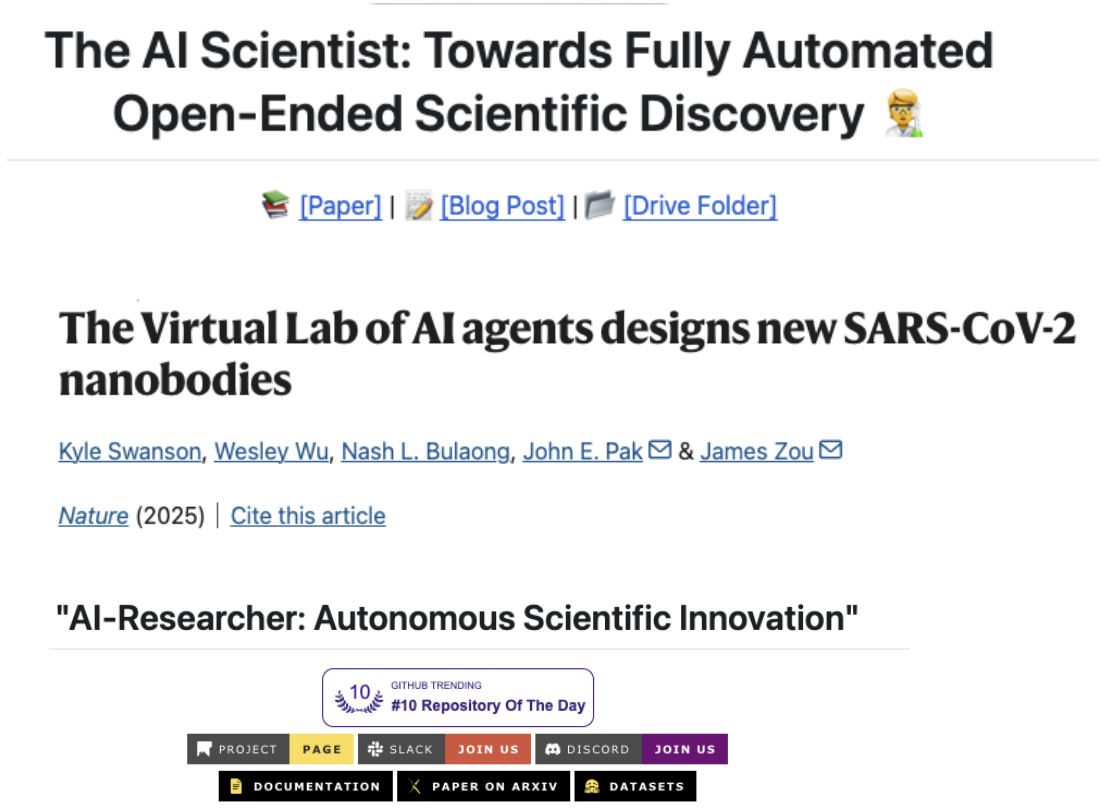
chenhao@uchicago.edu, @ChenhaoTan

https://ai-scientific-discovery.github.io/

AI will transform how science is done

But how?

The Explosion of AI Scientists



Posted on arXiv on Aug 12, 2024

Posted on bioRxiv on Nov 12, 2024

Posted on arXiv on May 24, 2025

Posted on arXiv on Nov 5, 2025

COMPOSITIONAL REGULARIZATION: UNEXPECTED OBSTACLES IN ENHANCING NEURAL NETWORK GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks excel in many tasks but often struggle with compositional generalization—the ability to understand and generate novel combinations of familiar components. This limitation hampers their performance on tasks requiring systematic reasoning beyond the training data. In this work, we introduce a training method that incorporates an explicit compositional regularization term into the loss function, aiming to encourage the network to develop compositional representations. Contrary to our expectations, our experiments on synthetic arithmetic expression datasets reveal that models trained with compositional regularization do not achieve significant improvements in generalization to unseen combinations compared to baseline models. Additionally, we find that increasing the complexity of expressions exacerbates the models' difficulties, regardless of compositional regularization. These findings highlight the challenges of enforcing compositional structures in neural networks and suggest that such regularization may not be sufficient to enhance compositional generalization.

1 INTRODUCTION

Compositional generalization refers to the ability to understand and produce novel combinations of known components, a fundamental aspect of human cognition (Ito et al., 2022). Despite the success of neural networks in various domains, they often struggle with compositional generalization, limiting their applicability in tasks requiring systematic reasoning beyond the training data (Qu et al., 2023; Klinger et al., 2020). Previous efforts to enhance compositional generalization have explored various approaches, including architectural modifications and training strategies (Finn et al., 2017; Lepori et al., 2023). One promising direction is the incorporation of regularization terms that encourage certain properties in the learned representations (Yin et al., 2023).

In this paper, we introduce a training method that incorporates an explicit *compositional regularization* term into the loss function. This regularization term is designed to penalize deviations from expected compositional structures in the network's internal representations, with the aim of encouraging the network to form compositional representations. We hypothesized that this approach would enhance the network's ability to generalize to unseen combinations. However, our experiments on synthetic arithmetic expression datasets show that the inclusion of compositional regularization does not lead to the expected improvements in generalization performance. In some cases, it even hinders the learning process. Furthermore, we observe that increasing the complexity of arithmetic expressions, such as using more operators or nesting, exacerbates the models' generalization difficulties regardless of the regularization. These unexpected results highlight the challenges of enforcing compositionality through regularization and suggest that this approach may not be straightforwardly effective.

In summary, we propose a compositional regularization term intended to enhance compositional generalization in neural networks, conduct extensive experiments to evaluate its impact, and analyze the unexpected outcomes, including the impact of operator complexity, discussing potential reasons why compositional regularization did not yield the anticipated benefits.

2 RELATED WORK

Compositional generalization in neural networks has been a topic of considerable research interest (Klinger et al., 2020). Ito et al. (2022) explored abstract representations to tackle this issue, emphasizing the importance of compositionality in achieving human-like reasoning. Yin et al. (2023) proposed consistency regularization training to enhance compositional generalization. Meta-learning approaches, such as Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017), have also been investigated to improve generalization capabilities. Lepori et al. (2023) studied structural compositionality in neural networks, suggesting that networks may implicitly learn to decompose complex tasks.

Our work differs by directly incorporating an explicit regularization term into the training objective to enforce compositional structures. Despite the theoretical appeal, our findings indicate that such regularization may not effectively enhance compositional generalization and that operator complexity plays a significant role in the models' performance limitations.

3 METHOD

Our goal is to enhance compositional generalization in neural networks by incorporating a compositional regularization term into the training loss. We focus on a simple yet illustrative task: evaluating arithmetic expressions involving basic operators.

3.1 MODEL ARCHITECTURE

We use an LSTM-based neural network (Goodfellow et al., 2016) to model the mapping from input expressions to their computed results. The model consists of an embedding layer, an LSTM layer, and a fully connected output layer.

3.2 COMPOSITIONAL REGULARIZATION

Let h_t be the hidden state at time t. We define the compositional regularization term as the mean squared difference between successive hidden states:

$$L_{\text{comp}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|h_{t+1} - h_t\|^2 \tag{1}$$

where T is the length of the input sequence.

This term penalizes large changes in hidden states between successive time steps, encouraging the model to form additive representations, which are a simple form of compositionality.

3.3 TRAINING OBJECTIVE

The total loss is the sum of the main loss (mean squared error between predicted and true results) and the compositional regularization term weighted by a hyperparameter λ:

$$L_{\text{total}} = L_{\text{main}} + \lambda L_{\text{comp}} \tag{2}$$

We experimented with different values of λ to assess its impact on compositional generalization.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We generated synthetic datasets of arithmetic expressions to evaluate compositional generalization. The datasets consist of expressions combining digits and operators (e.g., "3+4", "7*2"). We compared models trained with and without the compositional regularization term and performed several ablation studies to assess the impact of different hyperparameters, operator complexity, and architectural choices.

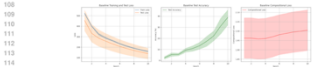


Figure 1: Baseline model performance over epochs. **Left**: Training and test loss decrease over epochs, indicating learning progress. **Middle**: Test accuracy increases, reaching approximately 84%. **Right**: Compositional loss remains steady, suggesting the model does not inherently develop compositional representations without regularization.

4.1.1 DATASETS

- **Training set**: 1,000 randomly generated expressions using a limited set of numbers and operators.
- **Test set**: 200 expressions not seen during training, including novel combinations of numbers and operators, as well as increased operator complexity.

4.1.2 IMPLEMENTATION DETAILS

- Models were trained for 30 epochs using the Adam optimizer and mean squared error loss.
- The compositional regularization term was weighted by $\lambda = 0.1$ unless otherwise specified.
- We evaluated model performance using test accuracy (percentage of correct predictions within a tolerance) and compositional loss.
- Experiments were repeated with different hyperparameters and operator complexities.

4.2 RESULTS

4.2.1 BASELINE PERFORMANCE

We first trained the baseline LSTM model without compositional regularization. Figure 1 shows the training and test loss, test accuracy, and compositional loss over epochs. As training progresses, both training and test loss decrease, and test accuracy increases, reaching approximately 84% accuracy. The compositional loss remains relatively steady, indicating that without regularization, the model does not inherently develop compositional representations.

4.2.2 IMPACT OF COMPOSITIONAL REGULARIZATION

We introduced the compositional regularization term with different weights λ and assessed its impact. Figure 2 illustrates the effects of varying λ on training loss, compositional loss, and final test accuracy. Higher values of λ led to a lower compositional loss but did not improve test accuracy. In some cases, the test accuracy decreased. This suggests that while compositional regularization encourages the learning of compositional representations as measured by the regularization term, it may interfere with the main learning objective by constraining the model's capacity to fit the training data.

4.2.3 IMPACT OF OPERATOR COMPLEXITY

We investigated how increasing the operator complexity of arithmetic expressions affects model performance. Figure 3 presents the training loss, validation loss, and final validation accuracy for expressions with varying numbers of operators. Our results show that as the complexity of the expressions increases, the models' ability to generalize diminishes significantly. Neither the baseline model nor the model with compositional regularization could handle expressions with higher operator complexity effectively. This finding emphasizes that compositional regularization alone may not address the challenges posed by complex compositional structures.

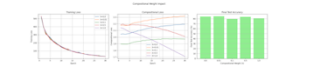


Figure 2: Impact of compositional weight λ on model performance. **Left**: Training loss over epochs for different λ. Higher λ values slightly increase training loss. **Middle**: Compositional loss decreases with higher λ, indicating the regularization term effectively enforces compositionality. **Right**: Final test accuracy does not improve with higher λ and may decrease, suggesting a trade-off between compositional regularization and the primary learning objective.

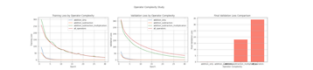


Figure 3: Model performance on expressions with varying operator complexity. **Left**: Training loss increases with operator complexity, indicating the models struggle to fit more complex data. **Middle**: Validation loss is higher for complex expressions, reflecting poor generalization. **Right**: Final validation accuracy decreases significantly as operator complexity increases, underscoring inherent limitations in handling complex compositional structures with compositional regularization alone.

5 CONCLUSION

In this work, we introduced a compositional regularization term with the intention of enhancing compositional generalization in neural networks. Our experiments on synthetic arithmetic expression datasets revealed that compositional regularization did not lead to the expected improvements in generalization performance. In some cases, it even hindered the learning process. Additionally, we found that increasing the complexity of arithmetic expressions exacerbates the models' generalization difficulties, highlighting inherent limitations.

These findings highlight the challenges of enforcing compositional structures in neural networks through regularization. Possible reasons for the lack of improvement include conflicts between the regularization term and the primary learning objective, which may cause the network to prioritize minimizing the compositional loss over fitting the data. Additionally, the measure of compositionality used in the regularization term may not align with the aspects of compositionality that are critical for generalization. The synthetic dataset may also not adequately capture the complexities of compositional generalization in real-world tasks, and increased operator complexity introduces additional challenges that compositional regularization alone cannot overcome.

For future work, we suggest exploring alternative regularization strategies, refining the definition of compositionality in the context of neural networks, and testing on more complex datasets. Investigating models that can inherently handle higher operator complexity, such as those with recursive or hierarchical structures, may also be beneficial. Our findings underscore the importance of rigorously evaluating proposed methods and openly reporting negative or inconclusive results to advance our understanding of the challenges in deep learning.

REFERENCES

Chelsea Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. pp. 1126–1135, 2017.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Takaya Ito, Tim Klinger, D. Schultz, J. Murray, Michael W. Cole, and Mattia Rigotti. Compositional generalization through abstract representations in human and artificial neural networks. 2022.

Tim Klinger, D. Adjodah, Vincent Marois, Joshua Joseph, M. Riemer, A. Pentland, and Murray Campbell. A study of compositional generalization in neural models. *ArXiv*, abs/2006.09437, 2020.

Michael A. Lepori, Thomas Serre, and Ellie Pavlick. Break it down: Evidence for structural compositionality in neural networks. *ArXiv*, abs/2301.10884, 2023.

Carolyn Qu, Rodrigo Nieto, • Mentor, and John Hewitt. Compositional generalization based on semantic interpretation: Where can neural networks improve? 2023.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. pp. 5998–6008, 2017.

Yongjing Yin, Jiali Zeng, Yafu Li, Fandong Meng, Jie Zhou, and Yue Zhang. Consistency regularization training for compositional generalization. pp. 1294–1308, 2023.

SUPPLEMENTARY MATERIAL

A EFFECT OF EMBEDDING DIMENSION

We explored the impact of different embedding dimensions on model performance. Figure 4 shows the training loss, compositional loss, and final test accuracy for embedding dimensions 16, 32, 64, and 128. Increasing the embedding dimension did not consistently improve test accuracy. While larger embedding dimensions provide the model with greater capacity, our results indicate that simply increasing model capacity is not sufficient to enhance compositional generalization in this context. This suggests that the bottleneck may lie in the model's ability to capture compositional structures rather than in its representational capacity.

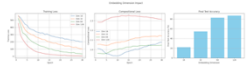


Figure 4: Effect of embedding dimension on model performance. **Left**: Training loss decreases similarly across embedding dimensions, indicating comparable learning progress. **Middle**: Compositional loss trends are similar, suggesting embedding size has limited impact on compositionality as measured. **Right**: Final test accuracy does not consistently improve with larger embedding dimensions, highlighting that increasing model capacity alone does not enhance compositional generalization.

B INTEGRATION OF ATTENTION MECHANISM

We compared the baseline model with an enhanced model that incorporates an attention mechanism (Vaswani et al., 2017). The attention mechanism is known to improve performance in various sequence-to-sequence tasks by allowing the model to focus on relevant parts of the input sequence.

B.1 EXPERIMENTAL SETUP

We modified the baseline LSTM model to include an attention layer after the LSTM outputs. The attention weights were calculated based on the hidden states, and a context vector was formed to aid in the final output prediction.

B.2 RESULTS

The attention model achieves a test accuracy similar to the baseline, as shown in Figure 5. While the attention mechanism slightly improved the training dynamics, it did not lead to significant improvements in generalization performance. This suggests that the challenges in compositional generalization are not primarily due to the model's ability to focus on relevant parts of the input sequence but may be related to deeper architectural limitations or the need for more sophisticated mechanisms to capture compositionality.

Figure 5: Comparison of baseline and attention models. **Left**: Training loss over epochs shows similar convergence for both models. **Middle**: Compositional loss remains comparable, indicating that attention does not significantly enhance compositional representations. **Right**: Final test accuracy is similar for both models, suggesting that the attention mechanism does not address the compositional generalization challenges.

C ADDITIONAL EXPERIMENTS

C.1 ABLATION STUDY ON COMPOSITIONAL WEIGHT

We conducted an ablation study on the compositional weight λ to further investigate its impact on model performance. Figures 6 and 7 show the training loss and final test accuracy for various values of λ. Higher λ values effectively reduce the compositional loss but adversely affect test accuracy. This reinforces the conclusion that emphasizing compositional regularization may conflict with the primary learning objective.

C.2 COMPARISON OF LSTM AND RNN ARCHITECTURES

We compared the performance of LSTM and simple RNN architectures to assess the influence of model choice on compositional generalization. Figure 8 illustrates the training loss and final test accuracy for both models. The LSTM model showed marginal improvements over the simple RNN, but both architectures struggled with compositional generalization, indicating that the limitations are not solely due to the recurrent unit type.

C.3 DROPOUT IMPACT

We investigated the impact of dropout on model performance. Figure 9 shows the final test accuracy for different dropout rates. We found that increasing the dropout rate did not lead to significant improvements in generalization, suggesting that regularization techniques like dropout may not address compositional generalization challenges. This indicates that standard regularization methods may not be sufficient to overcome the inherent difficulties in learning compositional structures.

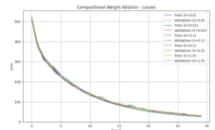


Figure 6: Training loss over epochs for different values of compositional weight λ. Increasing λ leads to slightly higher training loss, indicating potential interference with the primary learning objective.

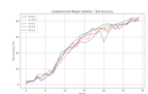


Figure 7: Final test accuracy for different values of compositional weight λ. Higher λ values do not improve test accuracy and may lead to decreased performance, suggesting a trade-off between compositional regularization and generalization.

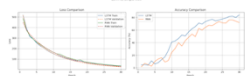


Figure 8: Comparison of LSTM and RNN architectures. **Left**: Training loss over epochs shows similar convergence patterns, with LSTM performing slightly better. **Right**: Final test accuracy is marginally higher for LSTM, but both models struggle with compositional generalization, suggesting that recurrent unit choice does not resolve the underlying challenges.

D HYPERPARAMETERS AND TRAINING DETAILS

We provide additional details on the hyperparameters and training procedures used in our experiments:

Figure 9: Final test accuracy for different dropout rates. Higher dropout rates did not enhance compositional generalization, indicating limited effectiveness of dropout in this context.

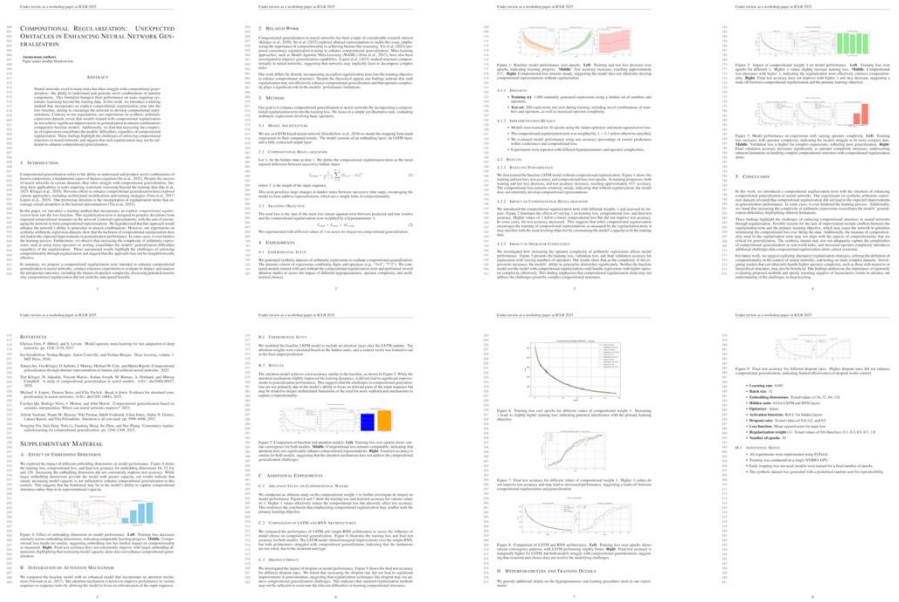
- **Learning rate**: 0.001
- **Batch size**: 32
- **Embedding dimensions**: Tested values of 16, 32, 64, 128
- **Hidden units**: 64 for LSTM and RNN layers
- **Optimizer**: Adam
- **Activation functions**: ReLU for hidden layers
- **Dropout rates**: Tested values of 0.0, 0.2, and 0.5
- **Loss function**: Mean squared error for main loss
- **Regularization weight** (λ): Tested values of 0.0 (baseline), 0.1, 0.3, 0.5, 0.7, 1.0
- **Number of epochs**: 30

D.1 ADDITIONAL NOTES

- All experiments were implemented using PyTorch.
- Training was conducted on a single NVIDIA GPU.
- Early stopping was not used; models were trained for a fixed number of epochs.
- The synthetic dataset was generated with a predefined random seed for reproducibility.

https://sakana.ai/ai-scientist-first-publication/

Funding
Research idea
Hypothesis
…



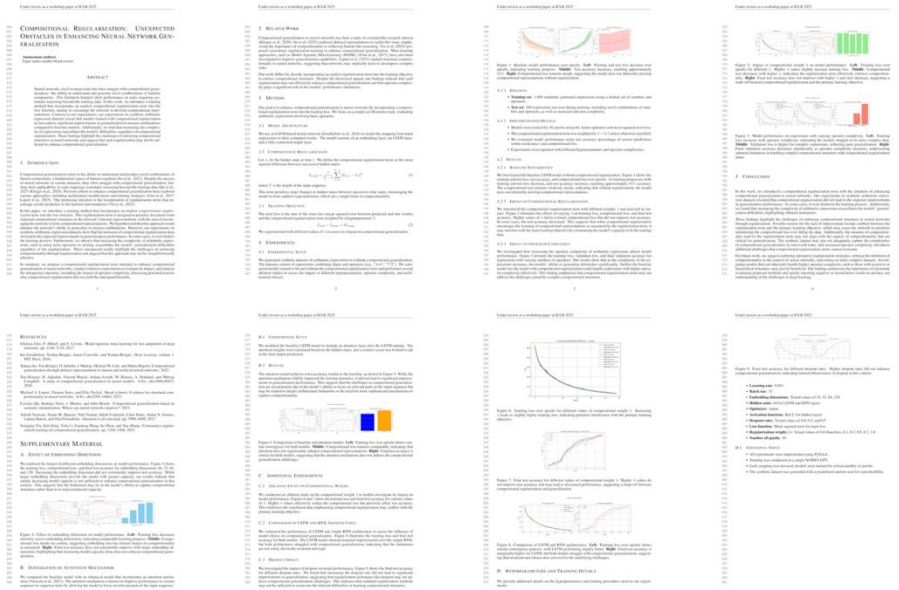
Attention
Coherence
Reproducibility
…

Selection

Production

Evaluation

Funding
Research idea
Hypothesis
…



Attention
Coherence
Reproducibility
…

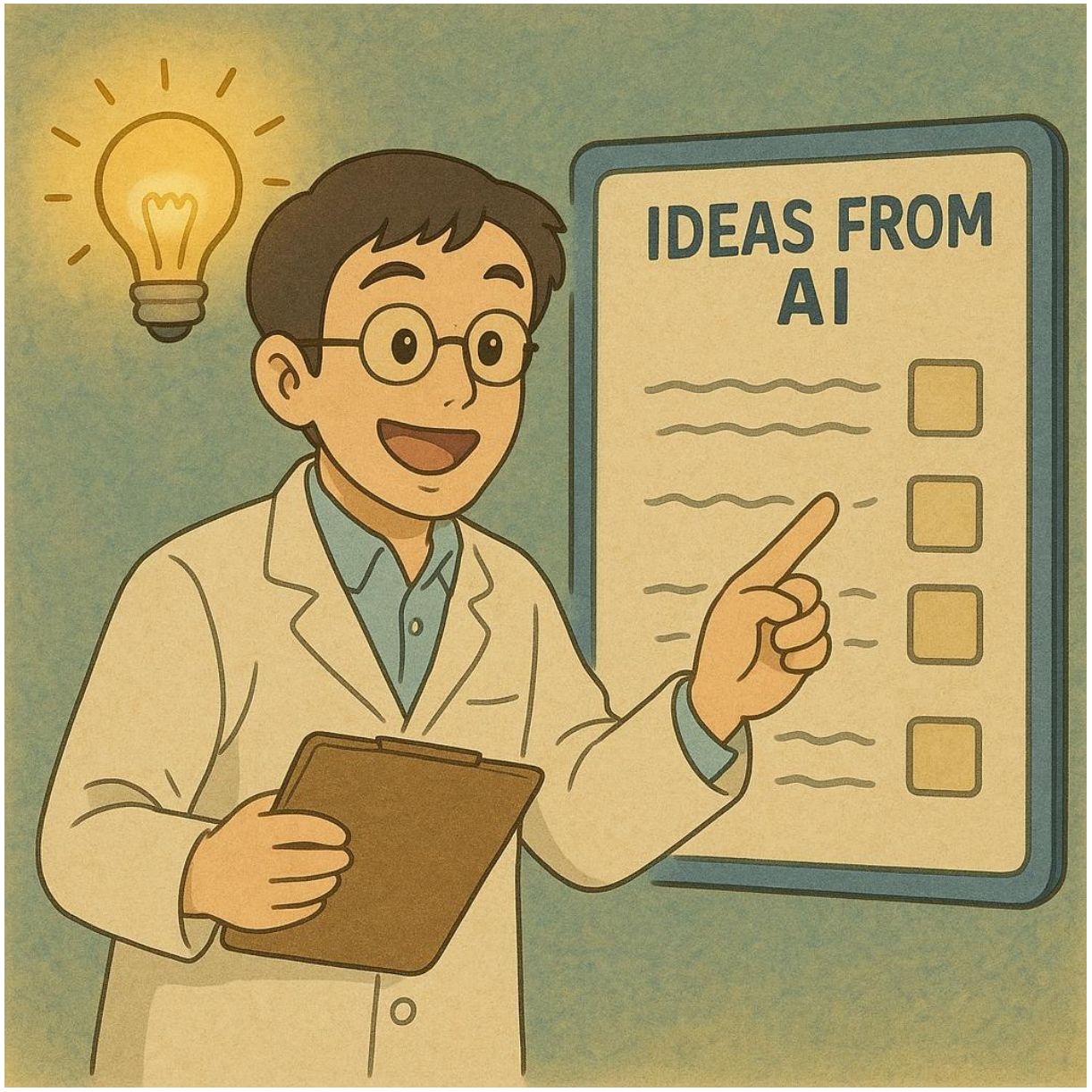
Selection

Production

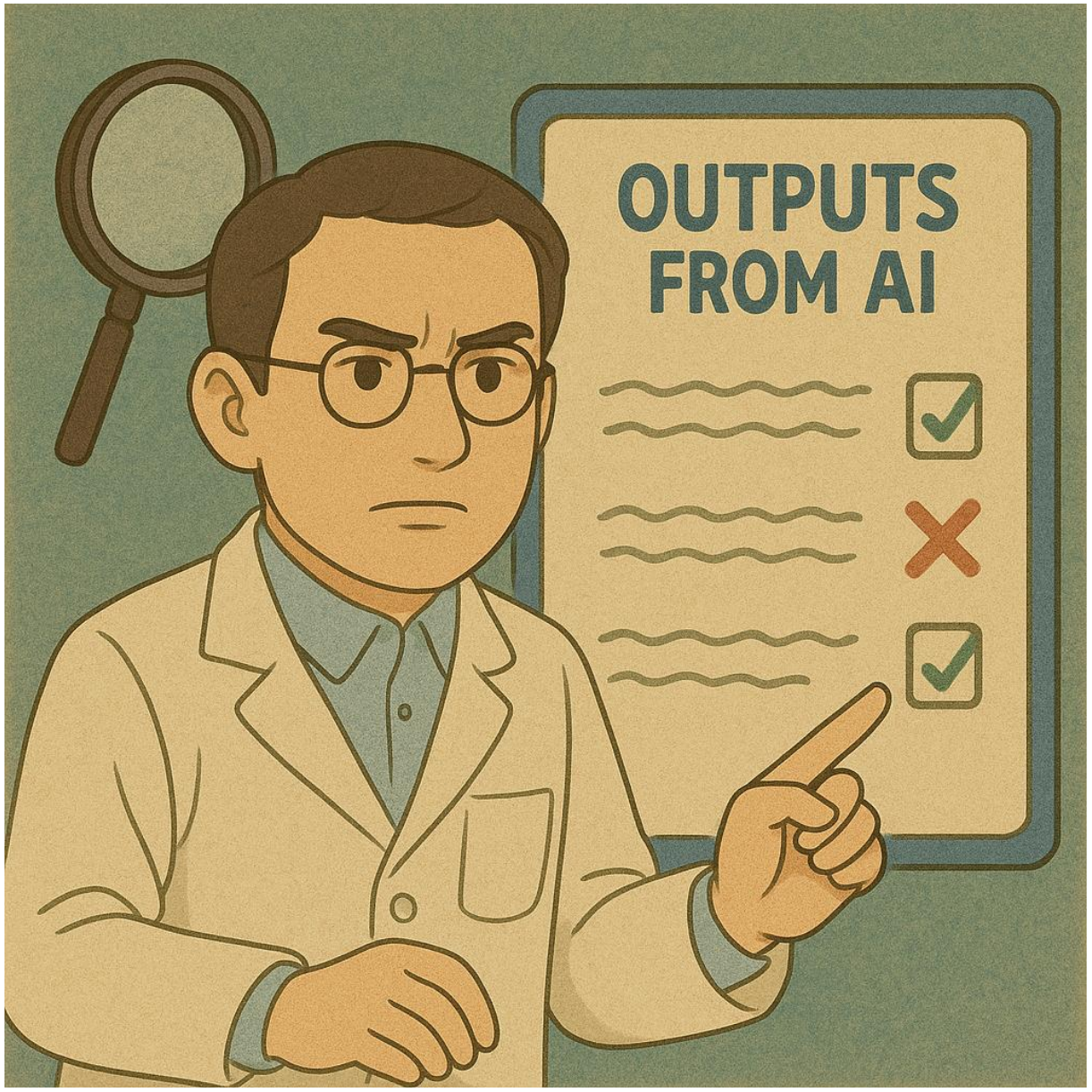
Evaluation

As AI takes over routine production tasks, the role of scientists will shift towards selection and evaluation

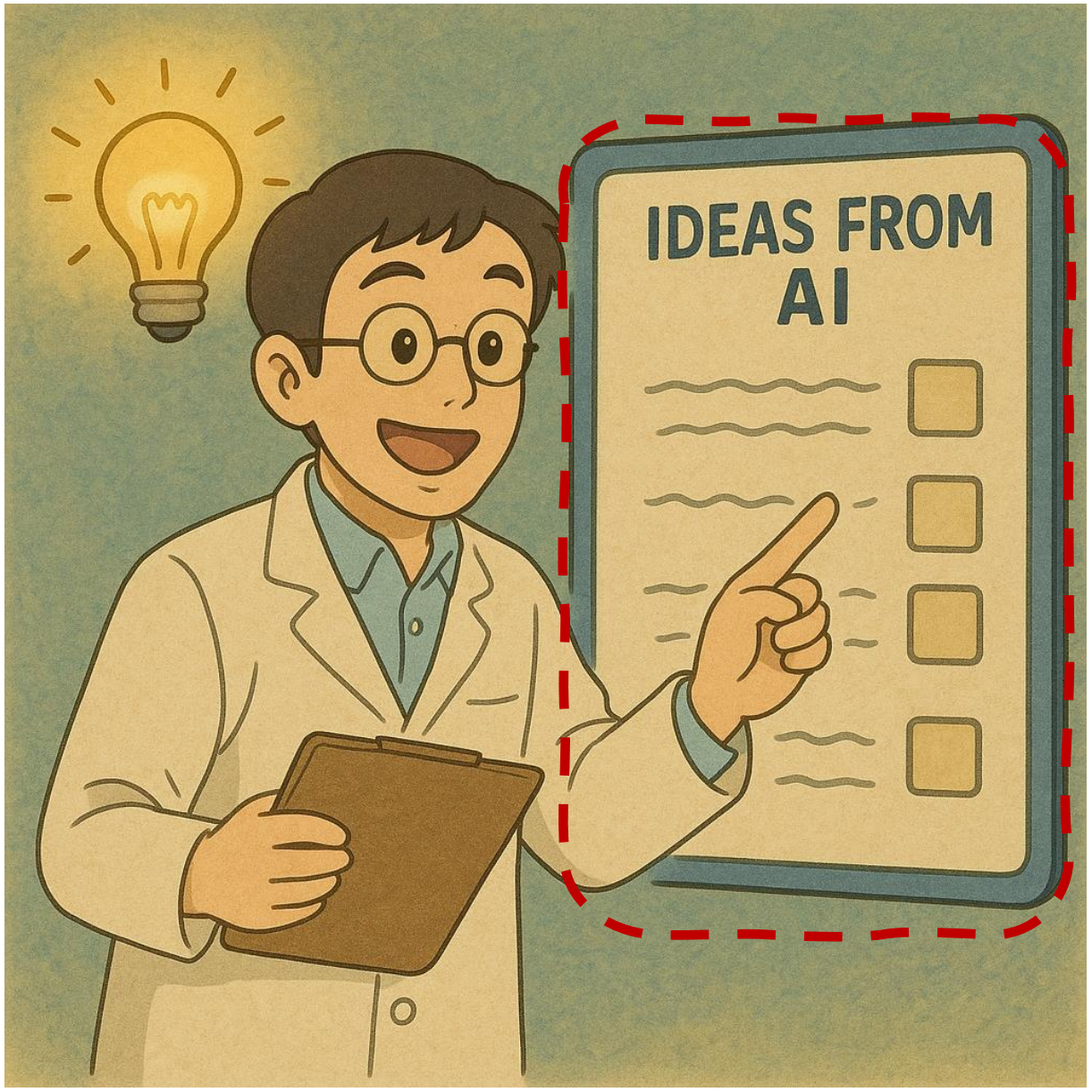
Selector



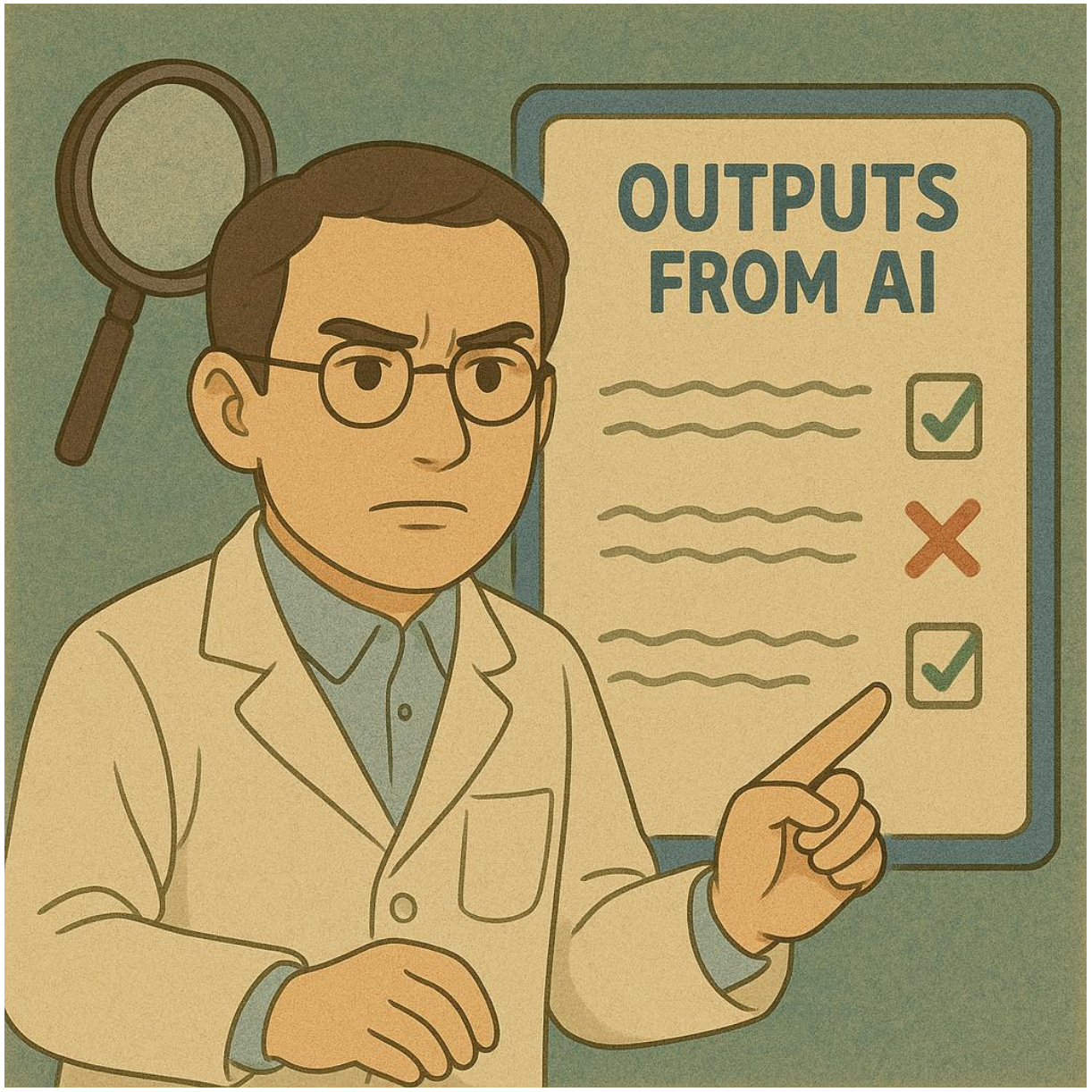
Evaluator



Selector

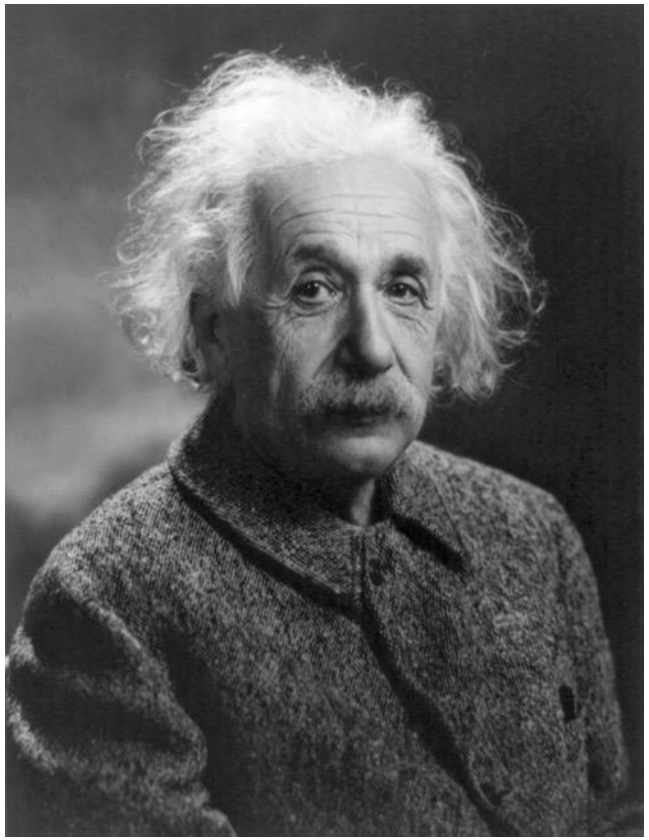


Evaluator

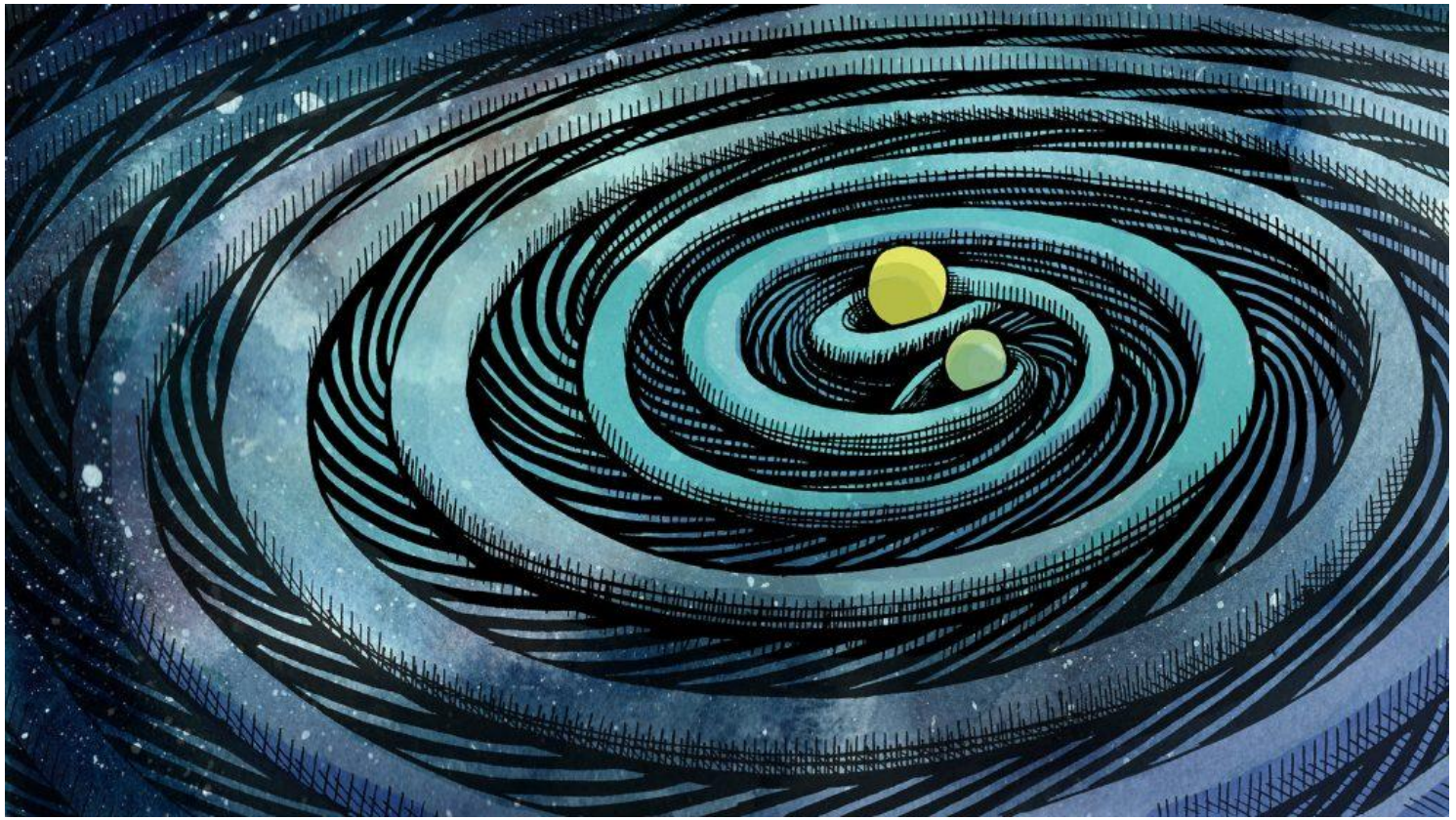


New theories (hypotheses) drive scientific progress

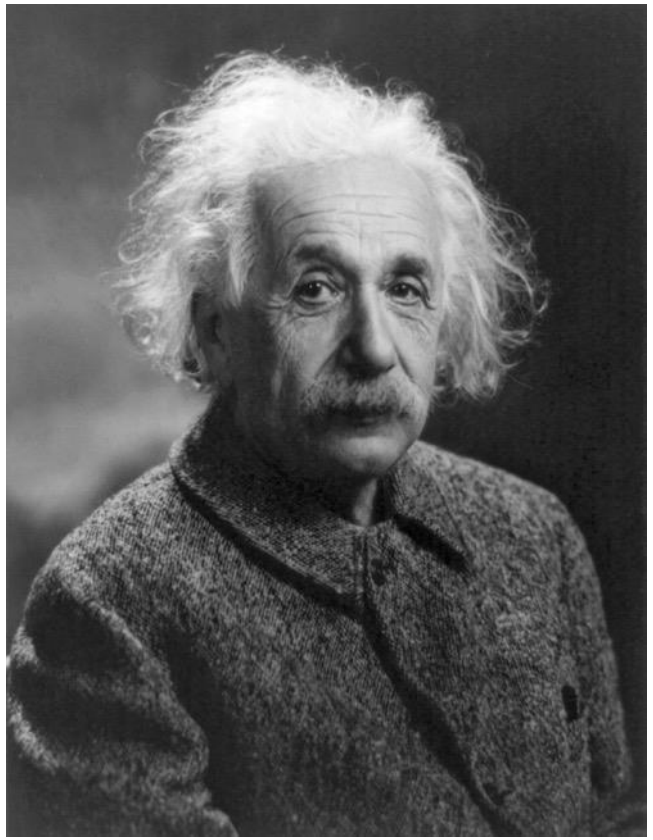
New theories (hypotheses) drive scientific progress



General theory of relativity

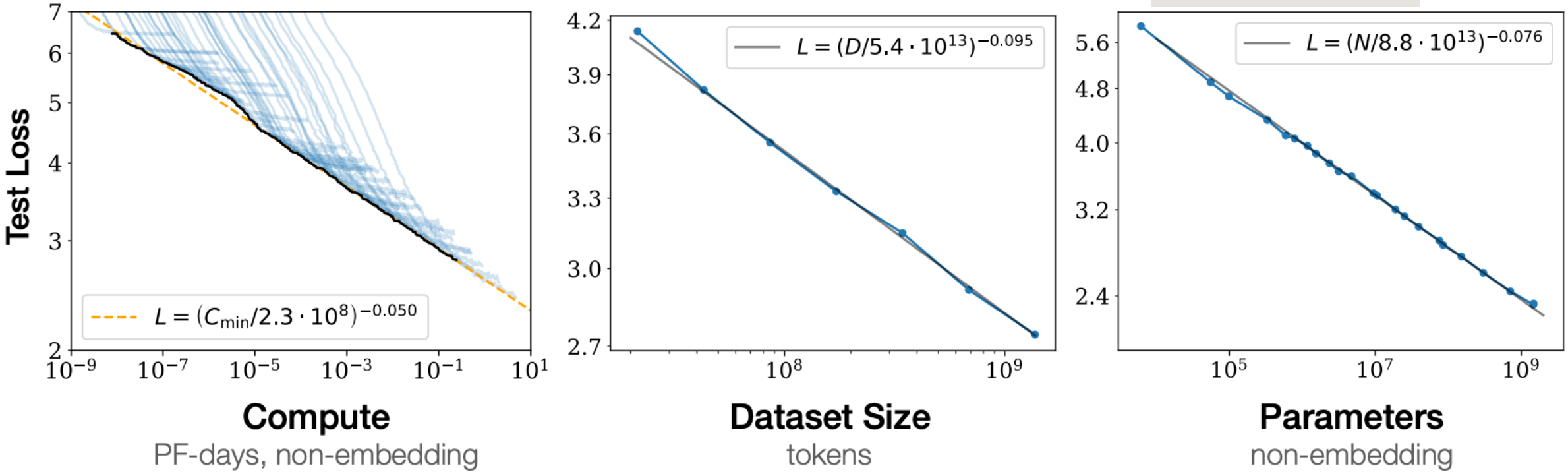


Discovery of gravitational waves



“It is the theory which decides what we can observe.”

New theories (hypotheses) drive scientific progress



Scaling Laws for Neural Language Models. Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei. 2020

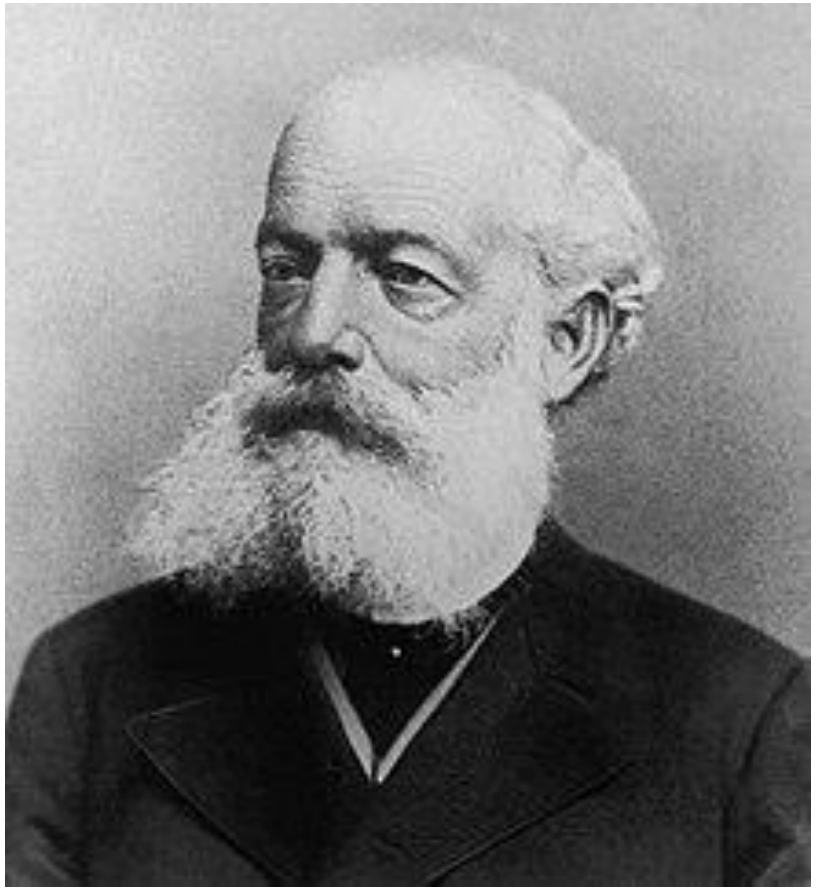
Despite the key role of hypotheses, the process of hypothesis generation has not been formalized, compared to hypothesis validation.

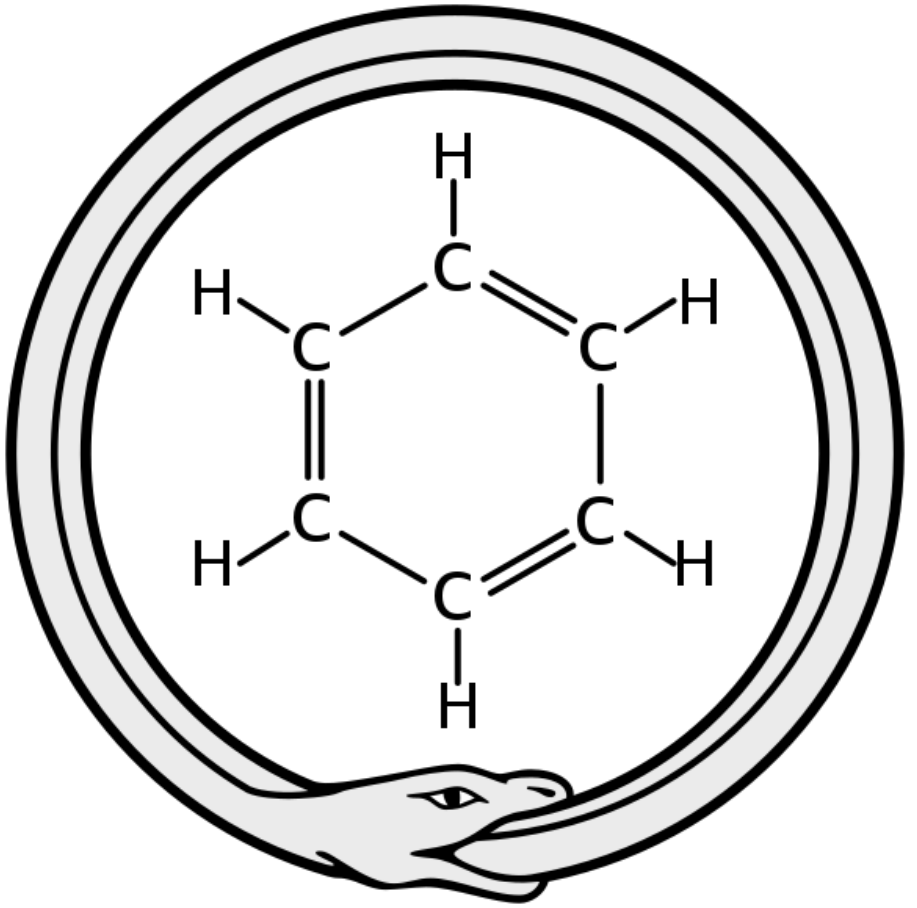
Where do hypotheses come from?

Where do hypotheses come from?

- Read literature
- Explore data
- Think

Where do hypotheses come from?

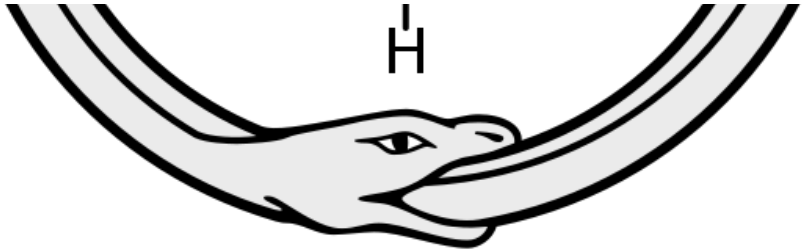




Where do hypotheses come from?

Creative cognitive processes in Kekulé's discovery of the structure of the benzene molecule

ALBERT ROTHENBERG
Harvard Medical School



Where do hypotheses come from?

Creative cognitive processes in Kekulé's discovery of the structure of the benzene molecule

ALBERT ROTHENBERG
Harvard Medical School

AI excels at "hallucination"

Where do hypotheses come from?

Creative cognitive processes in Kekulé's discovery of the structure of the benzene molecule

ALBERT ROTHENBERG
Harvard Medical School

AI excels at synthesizing and creating information

Build AI for hypothesis generation

Hypothesis Generation with Large Language Models.
Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, Chenhao Tan. NLP4Science at EMNLP 2024.

Literature Meets Data: A Synergistic Approach to Hypothesis Generation.
Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, Chenhao Tan. ACL 2025.





A concrete example: AIGC detection

The sun dipped low in the sky, casting a warm golden hue over the tranquil village of Eldergrove. The cobblestone streets were alive with the sounds of children laughing and adults chatting, but amid the bustle, Julian felt an expanding silence in his heart, an emptiness nurtured by years of questions, whispers, and the weight of uncertainty.

Example hypotheses

- AI-generated content uses more first-person pronouns.
- AI-generated content has consistent sentence structures.
- Human-written text has more informal languages and slangs.
- Human-written text has typos and grammatical errors.

Example hypotheses

- AI-generated content uses more first-person pronouns.
- AI-generated content has consistent sentence structures.
- Human-written text has more informal languages and slangs.
- Human-written text has typos and grammatical errors.

Hallucination is perfect for this goal!

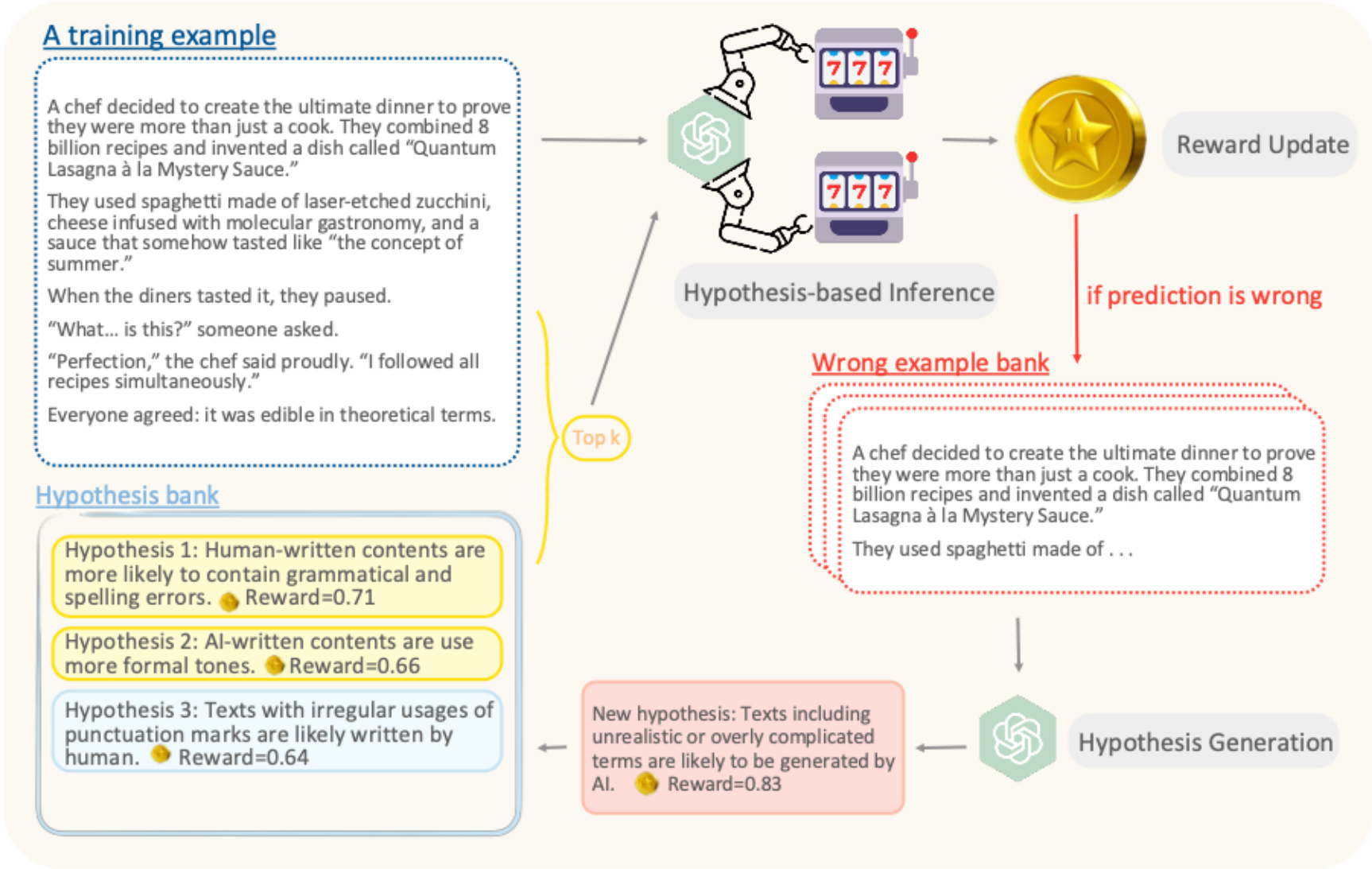
Formulating Hypothesis Generation

- Input:
  - A problem of interest (e.g., what characterizes AI-generated content)
  - Data (e.g., AI generated texts and human generated texts)
  - Related literature
- Output:
  - Natural language hypotheses that answer the problem of interest

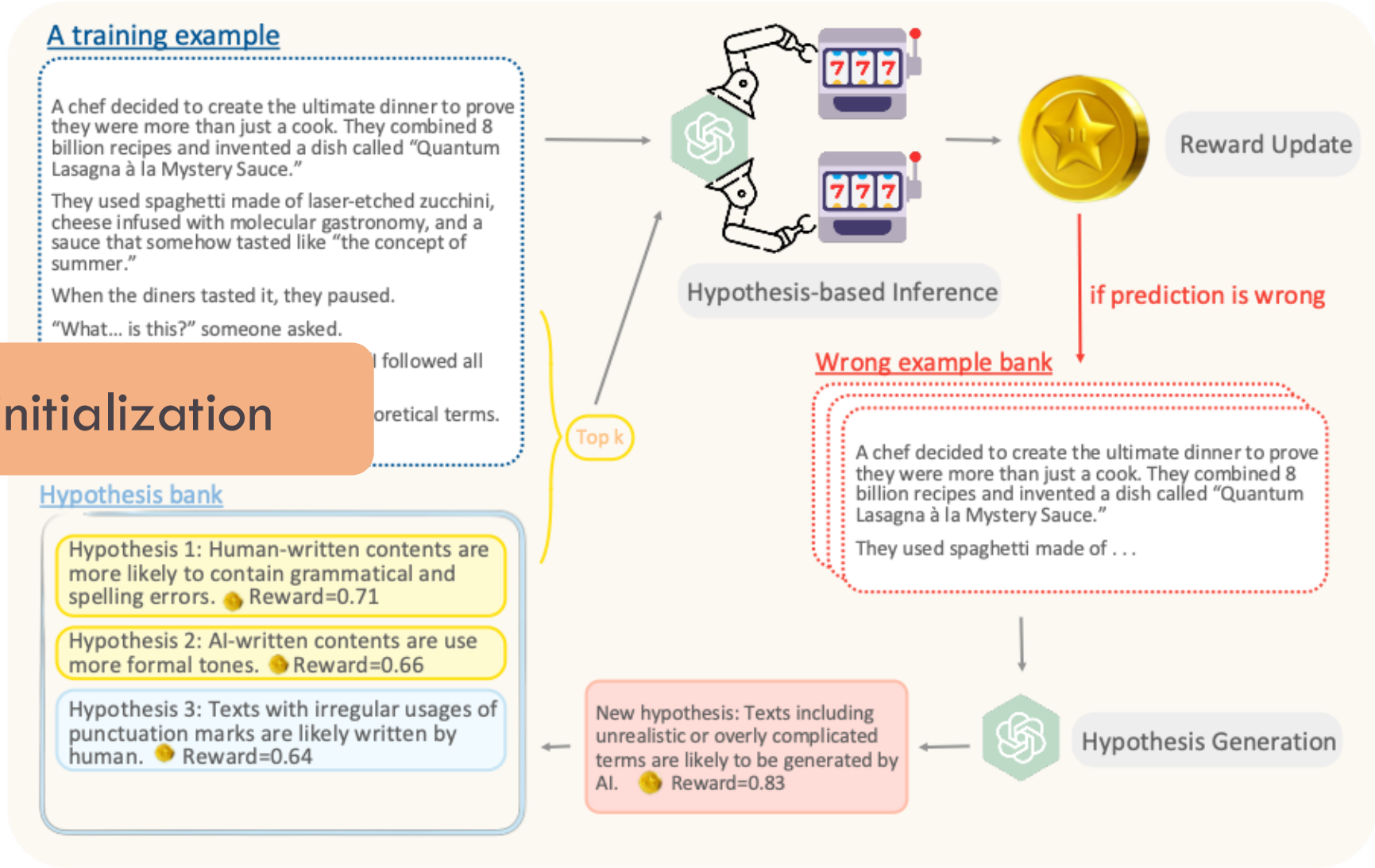
Two main approaches

- Data-driven: Look for patterns in data
  - Pro: Grounded in real data
  - Con: Overfitting
- Theory-driven: Building on existing theories
  - Pro: leveraging existing human knowledge
  - Con: limited by human knowledge

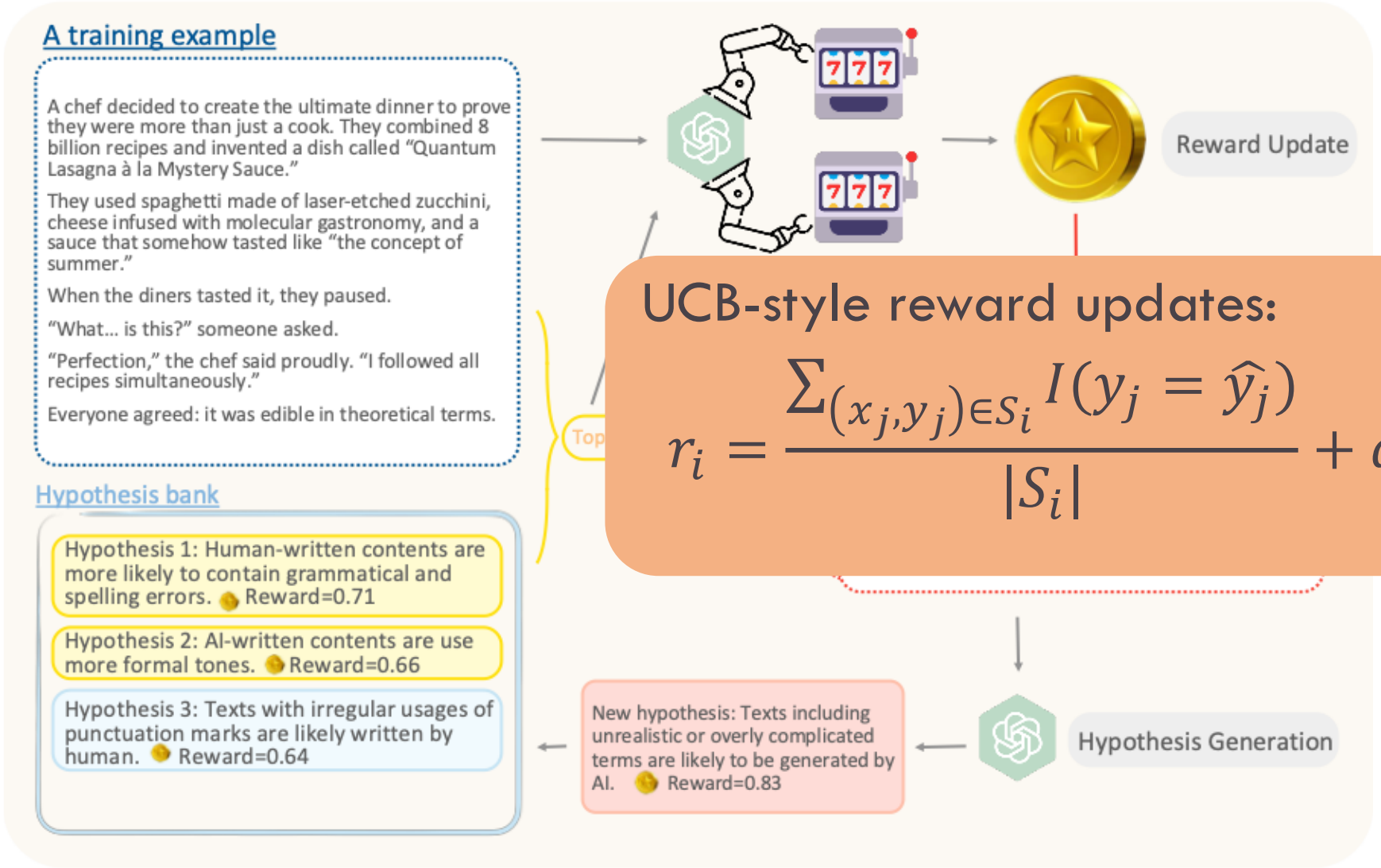
Hypogenic: A data-driven algorithm



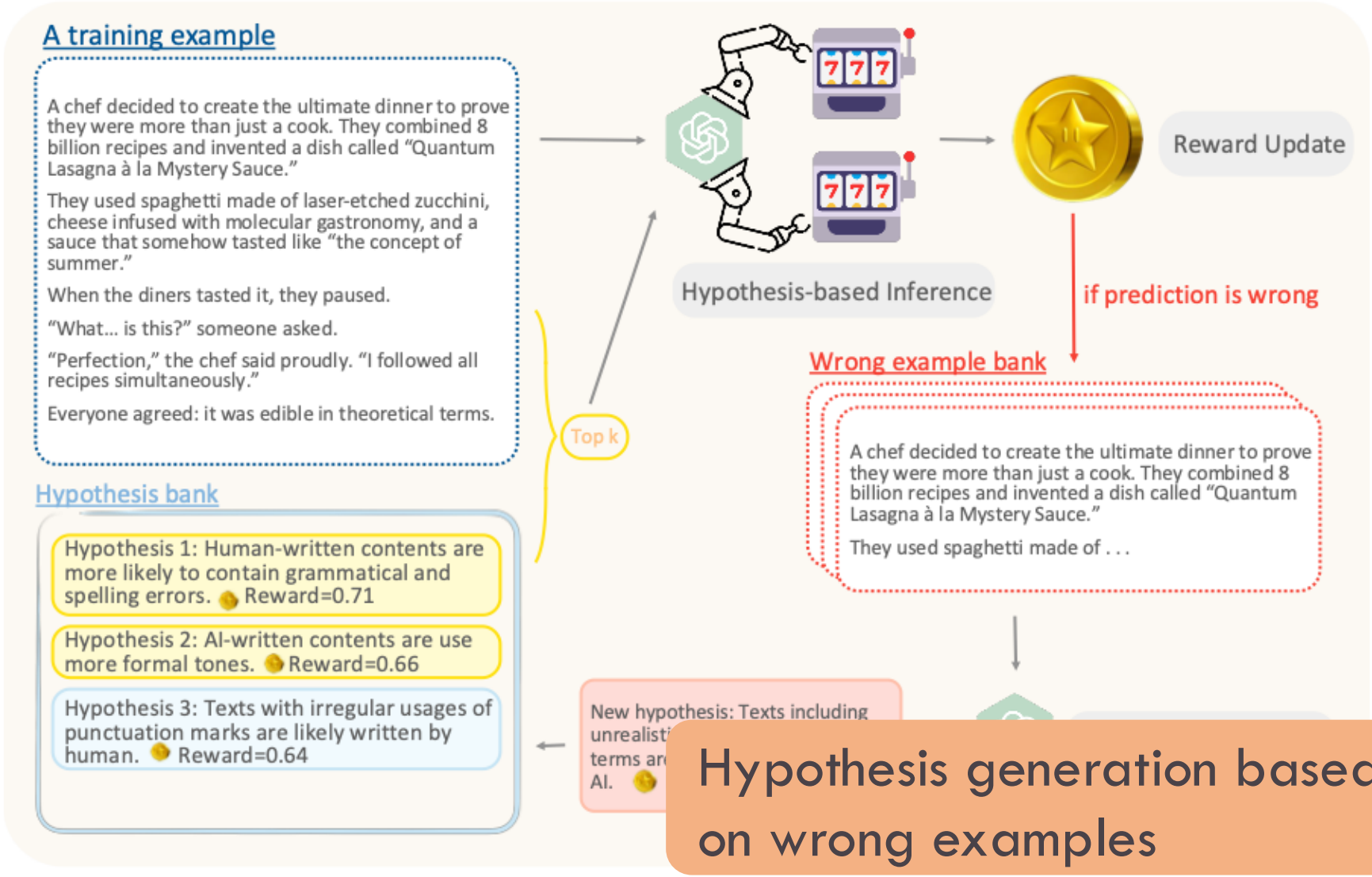
Hypogenic: A data-driven algorithm



Hypogenic: A data-driven algorithm

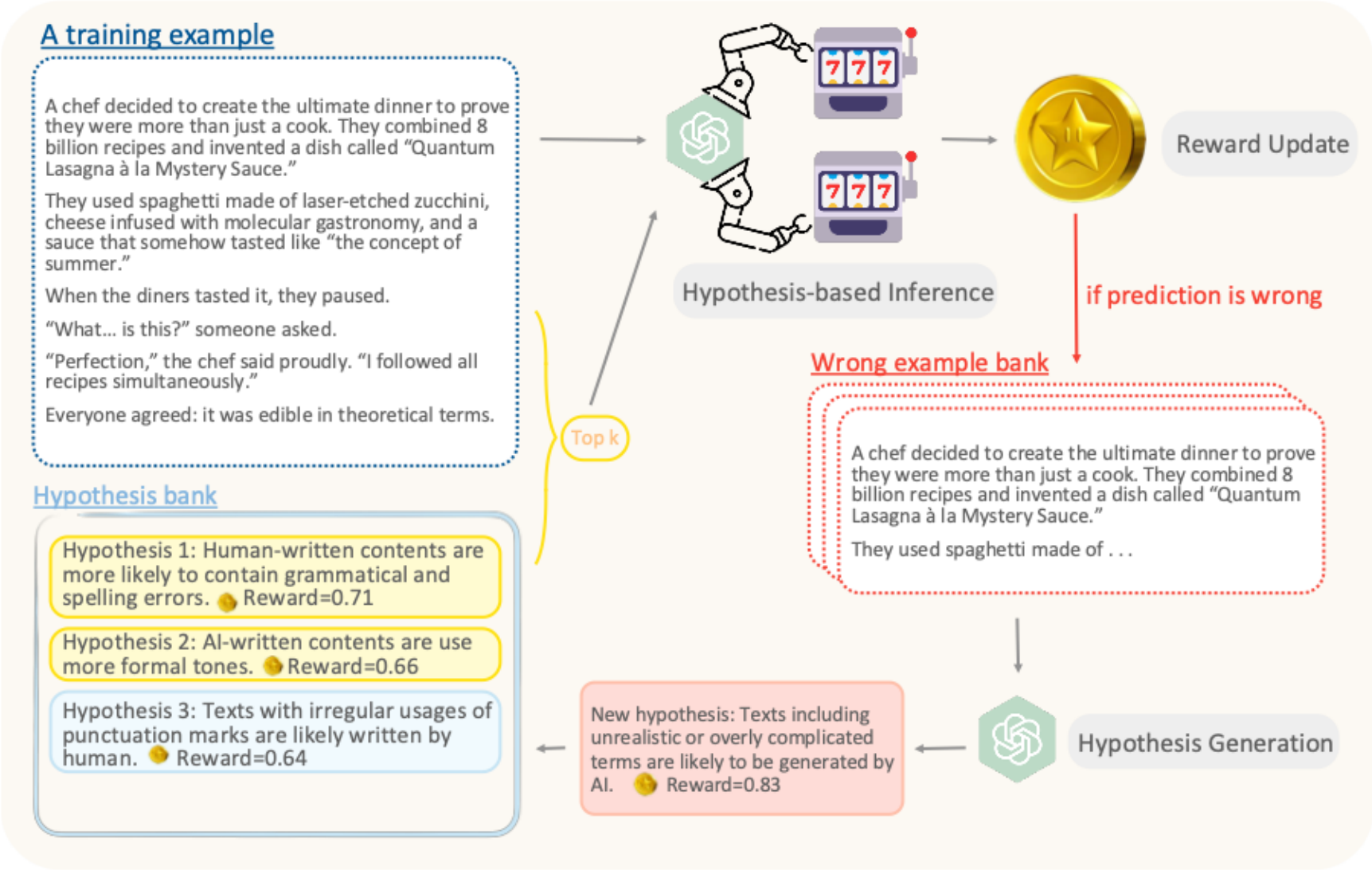


Hypogenic: A data-driven algorithm



Hypothesis generation based on wrong examples

Hypogenic: A data-driven algorithm



Use data labels to guide hallucinations

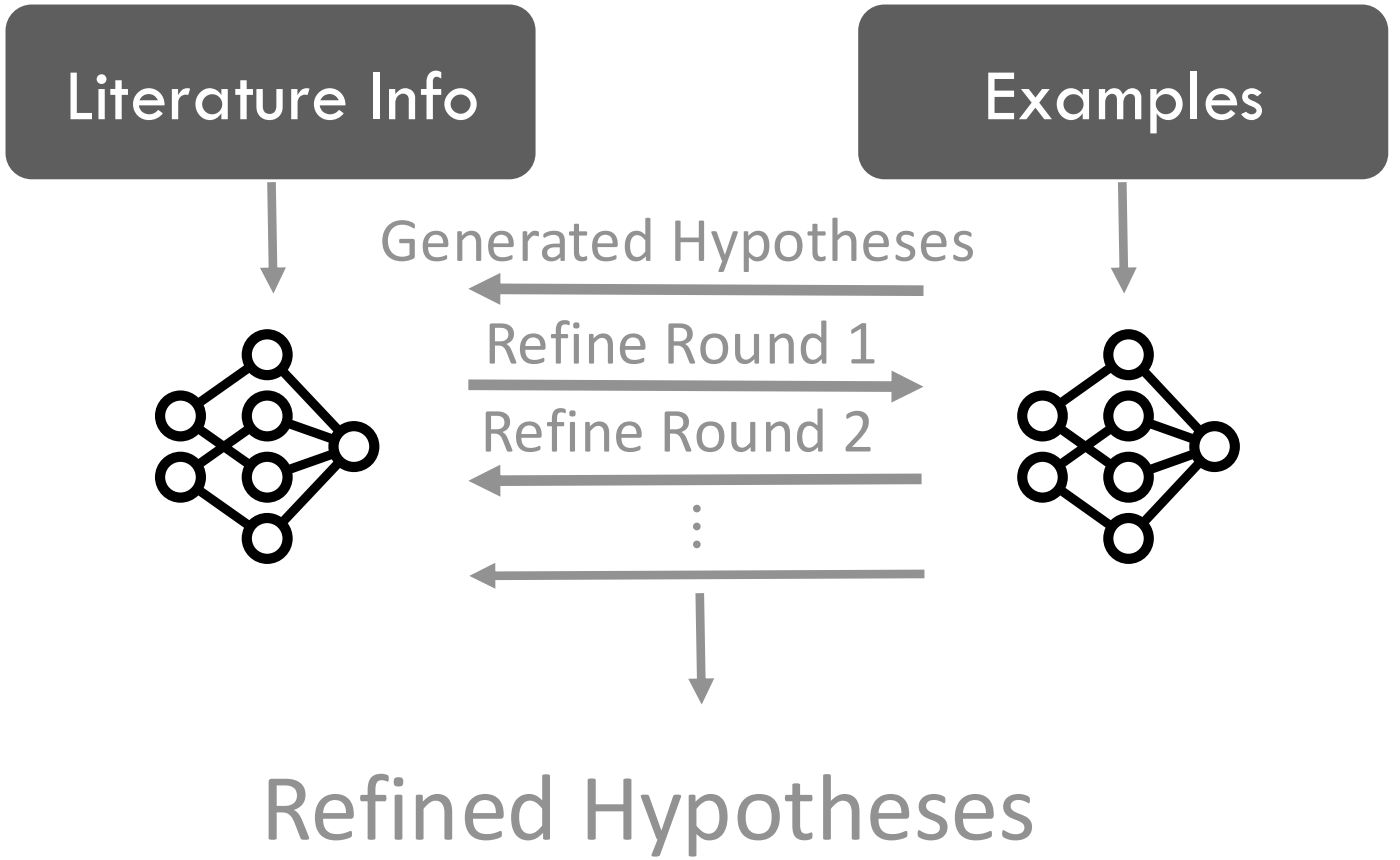
Literature-based hypothesis generation

Analogous to retrieval-augmented generation

- Search for relevant literature
- Summarize key findings of the retrieved literature
- Use key findings to generate hypotheses

Combining Hypogenic and Literature

- HypoRefine
- Literature + Hypogenic
- Literature + HypoRefine



Evaluation

- We can follow the recipe of supervised classification.
- However, what we care most about is

the quality of hypotheses:

  - Qualitative examination
  - Human evaluation
  - Cross-generalization

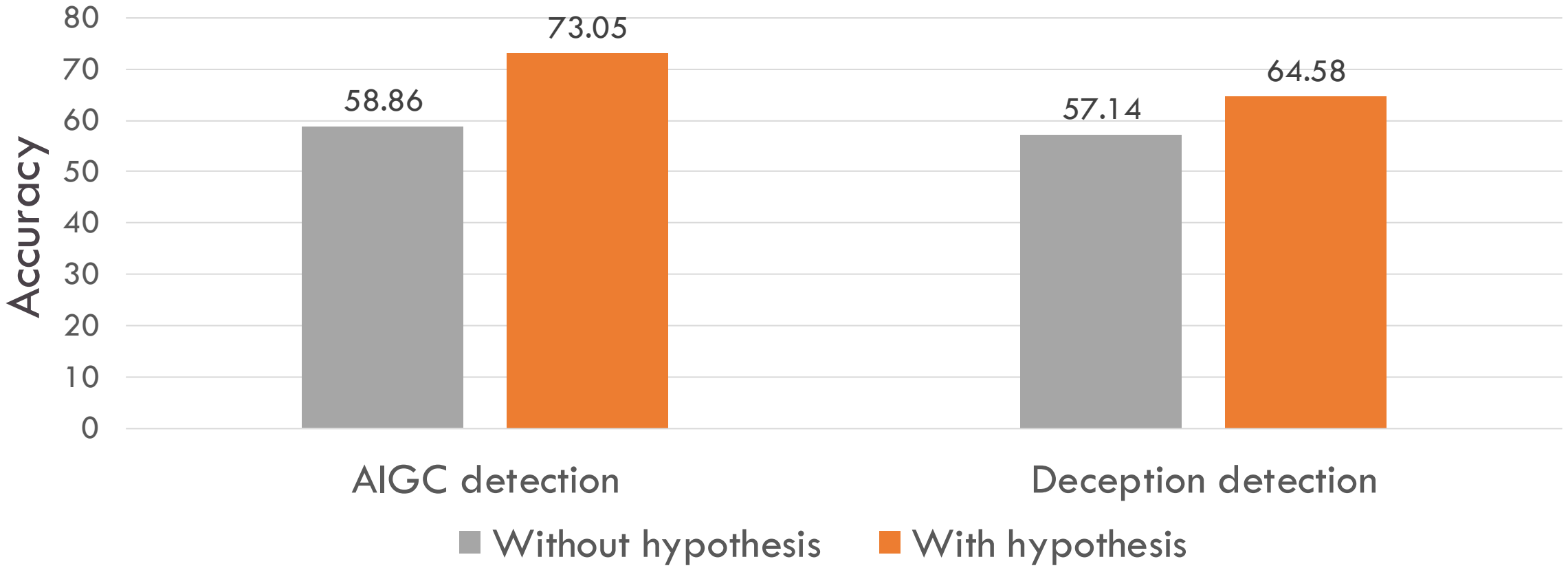
Example generated hypotheses for AIGC detection

- AI-generated texts tend to use more elaborate and descriptive language, including adjectives and adverbs, to create a sense of atmosphere and immersion. Human-written texts, on the other hand, tend to be more concise and straightforward in their language use.
- Human-written texts are more likely to contain errors or idiosyncrasies in grammar and punctuation, reflecting the natural imperfections of human writing, while AI-generated texts typically maintain a higher level of grammatical accuracy.
- Human-written texts tend to have more conversational tone and colloquial language, while AI-generated texts tend to be more formal and lack idiomatic expressions.

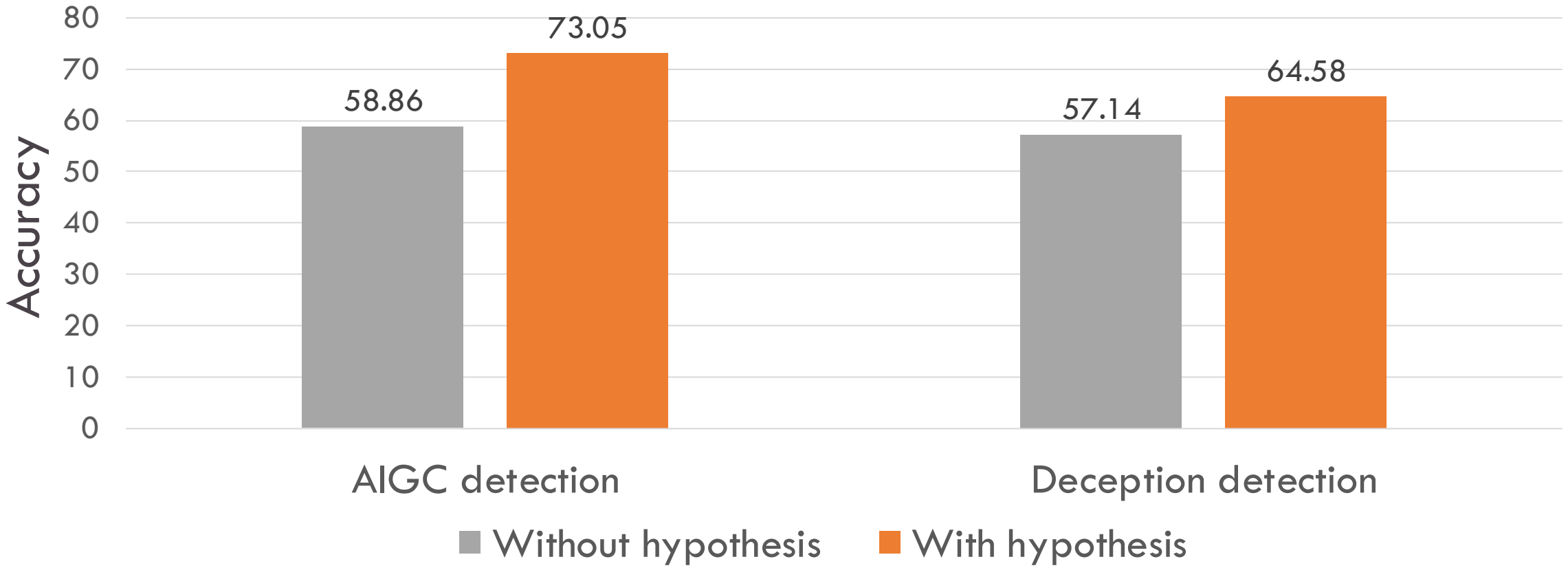
Example generated hypotheses for deception detection

- Reviews that present a balanced perspective by detailing both positive and negative experiences with specific examples (e.g., "the room was spacious and clean, but the noise from the street was disruptive at night") are more likely to be truthful, whereas reviews that express extreme sentiments without acknowledging any redeeming qualities (e.g., "everything was perfect" or "it was a total disaster") are more likely to be deceptive.
- Reviews that mention specific dates of stay or unique circumstances surrounding the visit (e.g., "We stayed during the busy Memorial Day weekend and faced long lines") are more likely to be truthful, while reviews that use vague temporal references (e.g., "I stayed recently") without concrete details are more likely to be deceptive, as they often lack the specificity that suggests a real and engaged experience.
- Reviews that provide detailed sensory descriptions of the hotel experience, such as the specific decor of the room, the quality of bedding, and the overall ambiance (e.g., "the room featured luxurious furnishings, high-thread-count sheets, and soft lighting that created a relaxing atmosphere") are more likely to be truthful, while reviews that use vague or overly simplistic descriptors (e.g., "the hotel was nice and comfortable") are more likely to be deceptive.

Generated hypotheses improve human decision-making



Generated hypotheses improve human decision-making



100% of the participants find the hypotheses to be helpful, and over 40% find them to be “Very helpful” or “Extremely helpful”.

Humans rate literature-based and data-driven hypotheses as distinct

- Case 1: Literature-only and Hypogenic generate different hypotheses

Literature-only: Deceptive reviews often contain a higher frequency of first-person singular pronouns, while truthful reviews may use these pronouns less frequently.

Hypogenic: Reviews that reference the reviewer's previous experiences with the hotel brand or similar hotels are more likely to be truthful, while reviews that do not provide any context or comparison to past experiences are more likely to be deceptive.

Humans rate literature-based and data-driven hypotheses as distinct

- Case 2: Literature-only and Hypogenic generate similar hypotheses

Literature-only: Truthful reviews often provide a balanced perspective, while deceptive reviews may seem overly promotional or biased towards a competitor.

Hypogenic: Reviews that express a balanced perspective, mentioning both positive and negative aspects of the stay, are more likely to be truthful, whereas reviews that are overly positive or negative without nuance tend to be deceptive.

Humans rate literature-based and data-driven hypotheses as distinct

- Case 2: Literature-only and Hypogenic generate similar hypotheses

HypoRefine: Reviews that present a balanced perspective by discussing both positive and negative aspects of the stay, particularly with specific examples (e.g., "The location was fantastic, but the air conditioning was broken"), are more likely to be truthful, while reviews that are excessively positive or negative without acknowledging any redeeming qualities (e.g., "This is the best hotel ever!" or "I will never stay here again!") tend to be more deceptive, as they may reflect an attempt to manipulate reader emotions rather than provide an honest assessment.

Automatic evaluation

- Five datasets:
  - Deception detection [Ott et al. 2013, Li et al. 2013]
  - GPTGC detection [Fan et al. 2018]
  - LlamaGC detection [Fan et al. 2018]
  - Persuasive argument detection [Pauli et al. 2024]
  - Mental stress detection (DREADDIT) [Turcan and McKeown 2019]
- We focus on out-of-distribution performance.
  - For example, LlamaGC is OOD for GPTGC.

Generated hypotheses outperform few-shot learning and other prompting approaches

Model	Methods	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
			No hypothesis			
GPT-4 MINI	Zero-shot	55.47	50.00	56.33	81.24	64.60
	Few-shot k=3	65.56	51.11	64.22	83.64	75.00
	Zero-shot generation	68.69	49.00	53.00	86.08	65.00
			Literature-based			
	LITERATURE-ONLY	59.22	49.00	54.00	78.80	67.68
	HYPERWRITE	61.63	49.67	52.67	82.36	68.76
	NOTEBOOKLM	53.03	49.33	51.67	68.96	62.28
			Data-driven			
	HYPOGENIC	75.22	81.67	68.56	82.20	76.56
			Literature + Data (This work)			
	HYPOREFINE	**77.78**	55.33	63.33	89.04	78.04
	Literature ∪ HYPOGENIC	72.41	**83.00**	**69.22**	**89.88**	78.20
	Literature ∪ HYPOREFINE	77.19	55.33	63.00	89.52	**79.24**

An average improvement of 11.92% over few-shot

Model	Methods	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
	No hypothesis					
	Zero-shot	55.47	50.00	56.33	81.24	64.60
	Few-shot k=3	65.56	51.11	64.22	83.64	75.00
	Zero-shot generation	68.69	49.00	53.00	86.08	65.00
	Literature-based					
GPT-4 MINI	LITERATURE-ONLY	59.22	49.00	54.00	78.80	67.68
	HYPERWRITE	61.63	49.67	52.67	82.36	68.76
	NOTEBOOKLM	53.03	49.33	51.67	68.96	62.28
	Data-driven					
	HYPOGENIC	75.22	81.67	68.56	82.20	76.56
	Literature + Data (This work)					
	HYPOREFINE	**77.78**	55.33	63.33	89.04	78.04
	Literature ∪ HYPOGENIC	72.41	**83.00**	**69.22**	**89.88**	78.20
	Literature ∪ HYPOREFINE	77.19	55.33	63.00	89.52	**79.24**

Commercial applications cannot do this task at all

Model	Methods	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
			No hypothesis			
	Zero-shot	55.47	50.00	56.33	81.24	64.60
	Few-shot k=3	65.56	51.11	64.22	83.64	75.00
	Zero-shot generation	68.69	49.00	53.00	86.08	65.00
			Literature-based			
GPT-4 MINI	LITERATURE-ONLY	59.22	49.00	54.00	78.80	67.68
	HYPERWRITE	61.63	49.67	52.67	82.36	68.76
	NOTEBOOKLM	53.03	49.33	51.67	68.96	62.28
			Data-driven			
	HYPOGENIC	75.22	81.67	68.56	82.20	76.56
			Literature + Data (This work)			
	HYPOREFINE	**77.78**	55.33	63.33	89.04	78.04
	Literature ∪ HYPOGENIC	72.41	**83.00**	**69.22**	**89.88**	78.20
	Literature ∪ HYPOREFINE	77.19	55.33	63.00	89.52	**79.24**

Literature can hurt hypothesis generation in the case of AIGC

Model	Methods	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
	No hypothesis					
	Zero-shot	55.47	50.00	56.33	81.24	64.60
	Few-shot k=3	65.56	51.11	64.22	83.64	75.00
	Zero-shot generation	68.69	49.00	53.00	86.08	65.00
	Literature-based					
GPT-4 MINI	LITERATURE-ONLY	59.22	49.00	54.00	78.80	67.68
	HYPERWRITE	61.63	49.67	52.67	82.36	68.76
	NOTEBOOKLM	53.03	49.33	51.67	68.96	62.28
	Data-driven					
	HYPOGENIC	75.22	81.67	68.56	82.20	76.56
	Literature + Data (This work)					
	HYPOREFINE	**77.78**	55.33	63.33	89.04	78.04
	Literature ∪ HYPOGENIC	72.41	**83.00**	**69.22**	**89.88**	78.20
	Literature ∪ HYPOREFINE	77.19	55.33	63.00	89.52	**79.24**

Generated hypotheses can be effectively transferred to a different model

Generation Model	Inference Model	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
		OOD Accuracy	OOD Accuracy	OOD Accuracy	OOD Accuracy	OOD Accuracy
GPT-4-MINI	GPT-4-MINI	77.78	83.00	69.22	89.88	79.24
	LLAMA-70B-I	72.53 (↓5.25)	71.67 (↓11.33)	76.33 (↑7.11)	86.88 (↓3.00)	72.36 (↓6.88)
LLAMA-70B-I	LLAMA-70B-I	73.72	81.33	78.67	88.76	78.92
	GPT-4-MINI	70.31 (↓3.41)	57.00 (↓24.33)	74.67 (↓4.00)	89.36 (↑0.60)	77.28 (↓1.64)

Generated hypotheses can be effectively transferred to a different model

Generation Model	Inference Model	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
		OOD Accuracy	OOD Accuracy	OOD Accuracy	OOD Accuracy	OOD Accuracy
GPT-4-MINI	GPT-4-MINI	77.78	83.00	69.22	89.88	79.24
	LLAMA-70B-I	72.53 (↓5.25)	71.67 (↓11.33)	76.33 (↑7.11)	86.88 (↓3.00)	72.36 (↓6.88)
LLAMA-70B-I	LLAMA-70B-I	73.72	81.33	78.67	88.76	78.92
	GPT-4-MINI	70.31 (↓3.41)	57.00 (↓24.33)	74.67 (↓4.00)	89.36 (↑0.60)	77.28 (↓1.64)

Our methods still outperform the few-shot inference baseline by 3.76%.

AI will drive future hypothesis generation

Next Steps

HypoBench: Towards Systematic and Principled Benchmarking for Hypothesis Generation.
Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, Chenhao Tan.

Heuristic-Based Ideation for Guiding LLMs Toward Structured Creativity.
Xiao Liu, Haokun Liu, Chenhao Tan.

Principled and systematic benchmarking of hypothesis generation

- Concept clarification
  - ***Hypothesis generation*** aims to generating natural language theories/explanations about observed phenomena
  - ***Research ideation*** aims to generate new research directions, primiarily from existing scientific literature
- Lack of access to groundtruth hypotheses
  - Synthetic datasets with groundtruth hypotheses

12 Domains;
194 Datasets

Real

- Deceptive Reviews Detection
- AI-generated Content Detection
- Persuasive Argument Prediction
- Paper Citations
- Mental Stress Detection
- News Headline Engagement
- Retweets

Synthetic

- Presidential Election
- Personality Prediction
- Marine Ecosystem
- College Admission
- Shoe Sales

Difficulty Control

Using college admission as an example

Original hypothesis:
Students with an A in Math will be admitted, otherwise rejected.

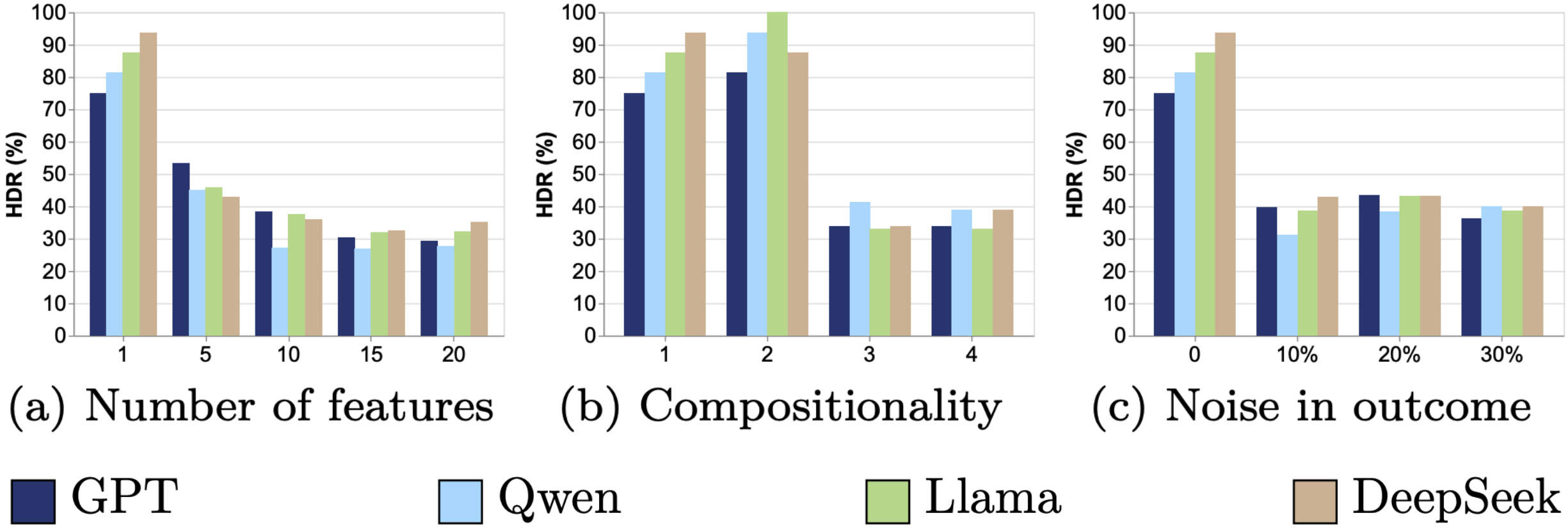
Noise:
Students with an A in Math will be admitted, otherwise rejected.
(10% chance the label is flipped)

Feature interaction:
Students with A in Math and at least one publication will be admitted, otherwise rejected.

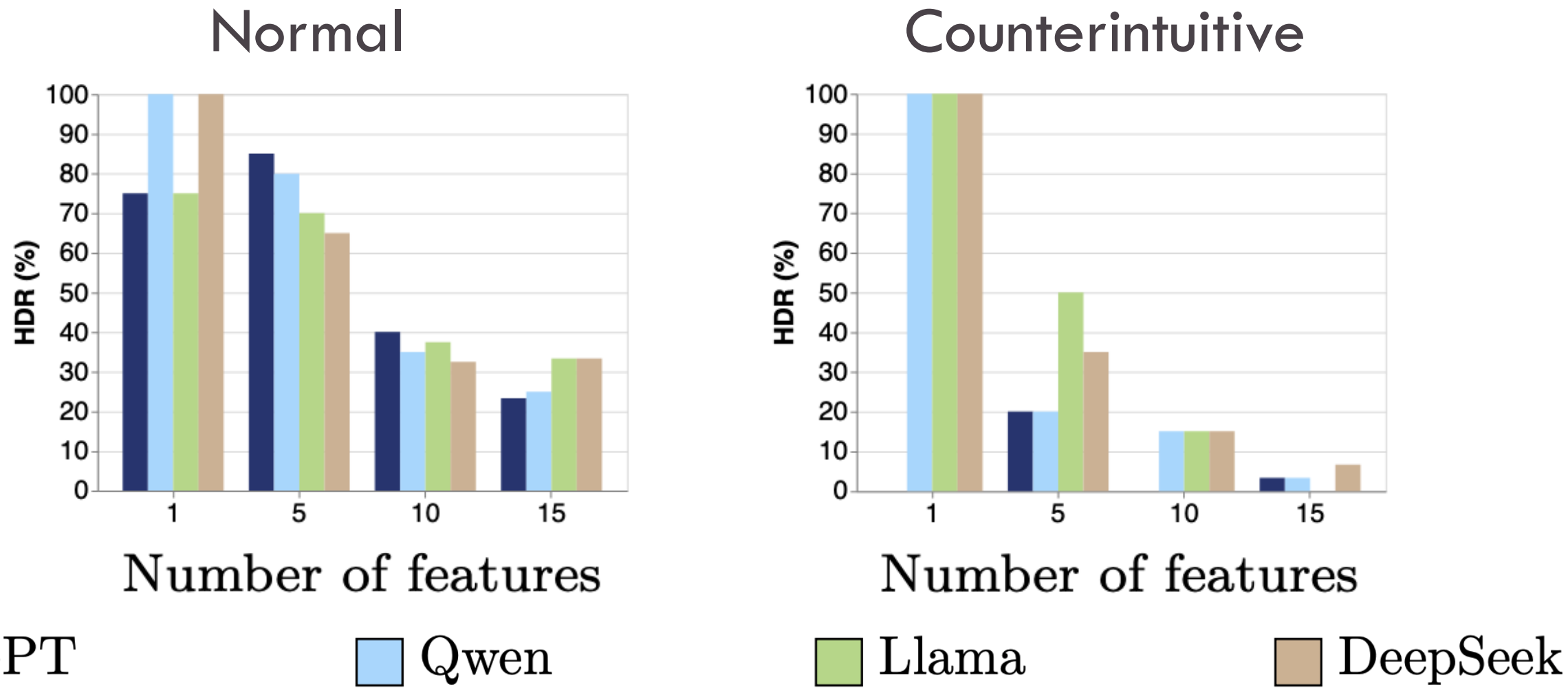
Increase number of features:
Students with an A in Math will be admitted.
Students with more than 2 strong activities will be admitted.
Students will be admitted if they are legacy students.

Add distractor features:
Students with an A in Math will be admitted, otherwise rejected.
Distractor Features:
- Extracurricular activities
- Legacy

Hypothesis discovery rate substantially drops, even to below 30% sometimes.



Model performance drops further for counterintuitive hypotheses



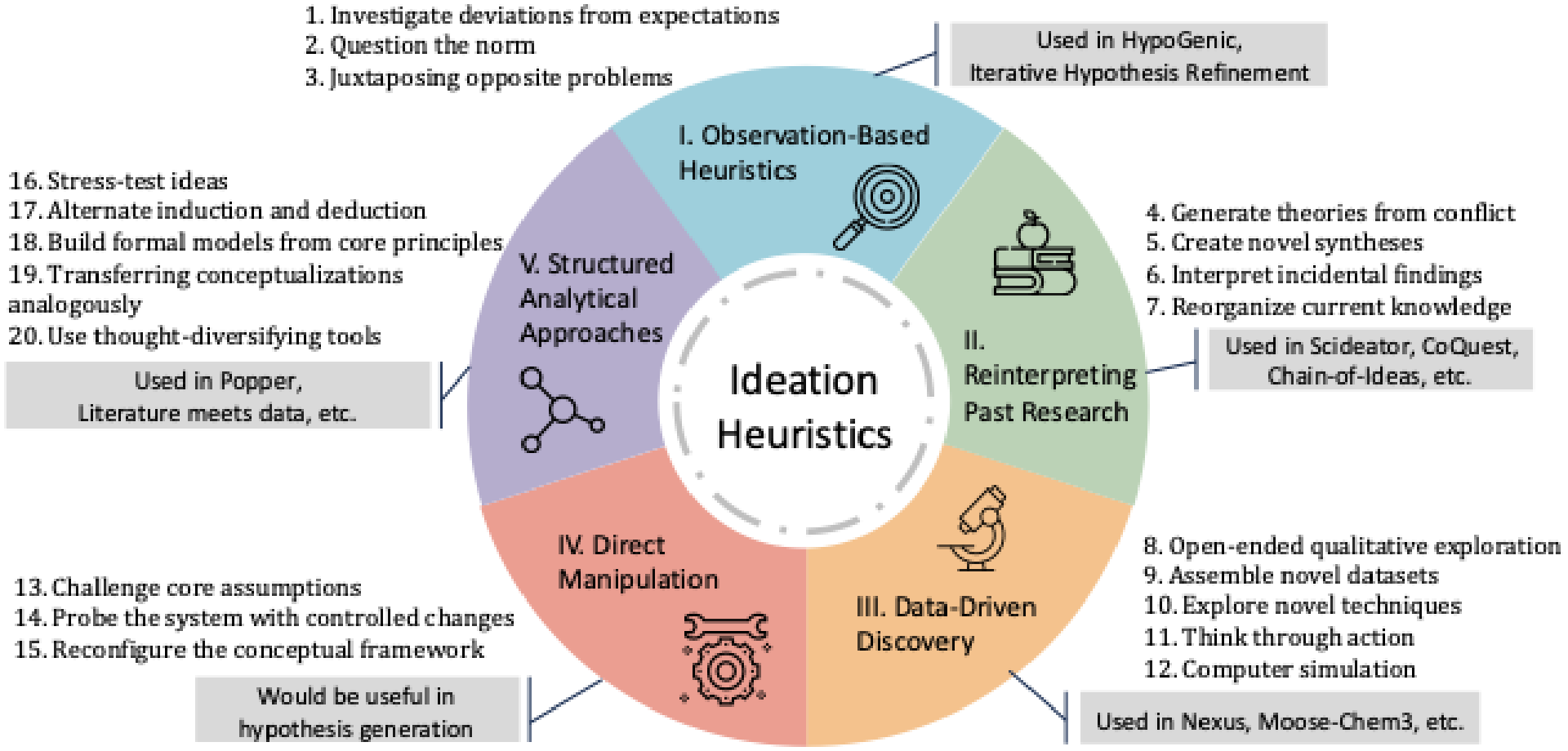
Formalizing Research Ideation and Hypothesis Generation

CREATIVE HYPOTHESIS GENERATING IN PSYCHOLOGY: Some Useful Heuristics

William J. McGuire

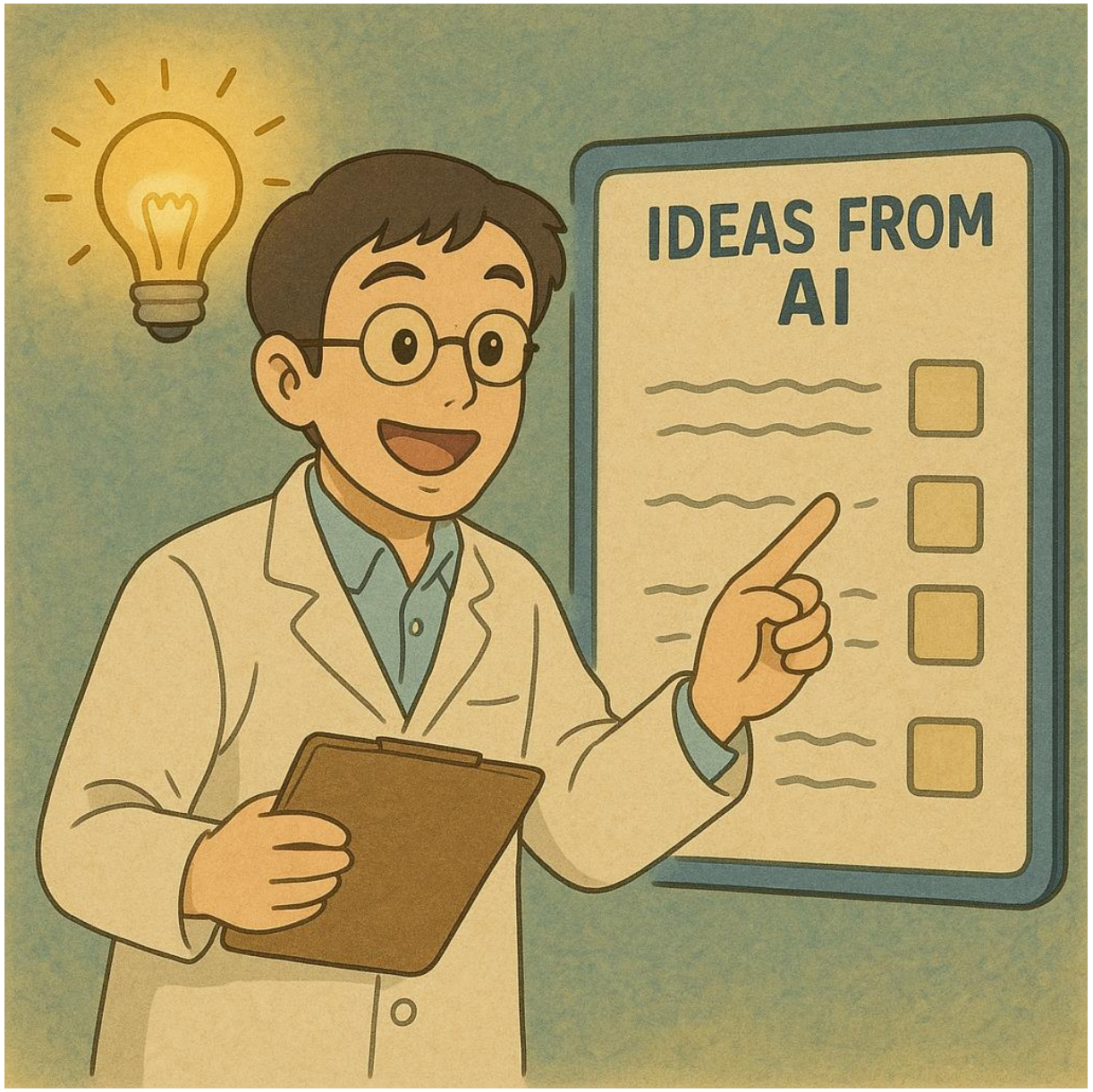
Yale University, Department of Psychology, P.O. Box 208205, 2 Hillhouse Avenue, New Haven, Connecticut 06520-8205

Formalizing Research Ideation and Hypothesis Generation

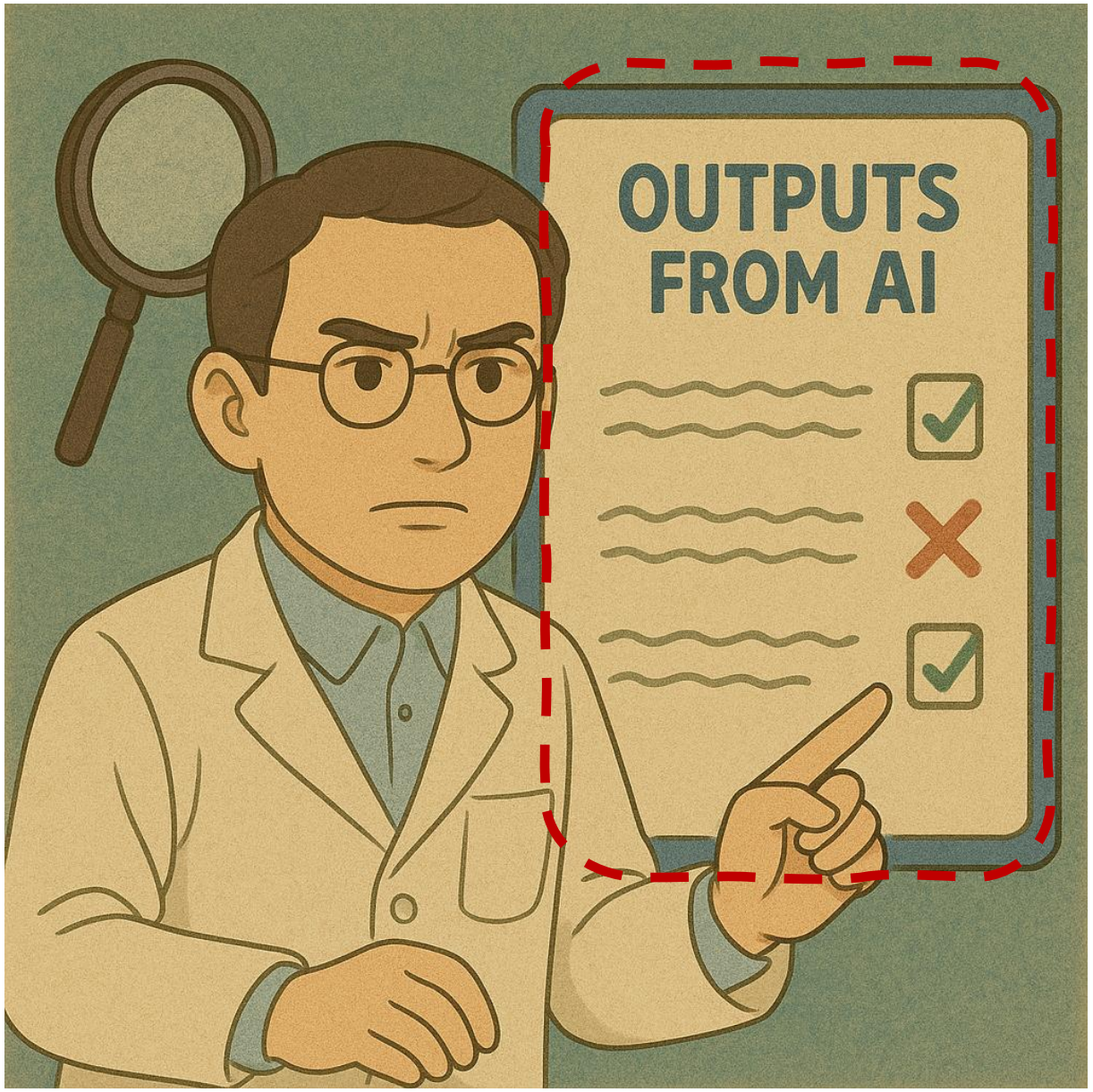


https://github.com/heuristic-ideation/heuristic-ideation-skill

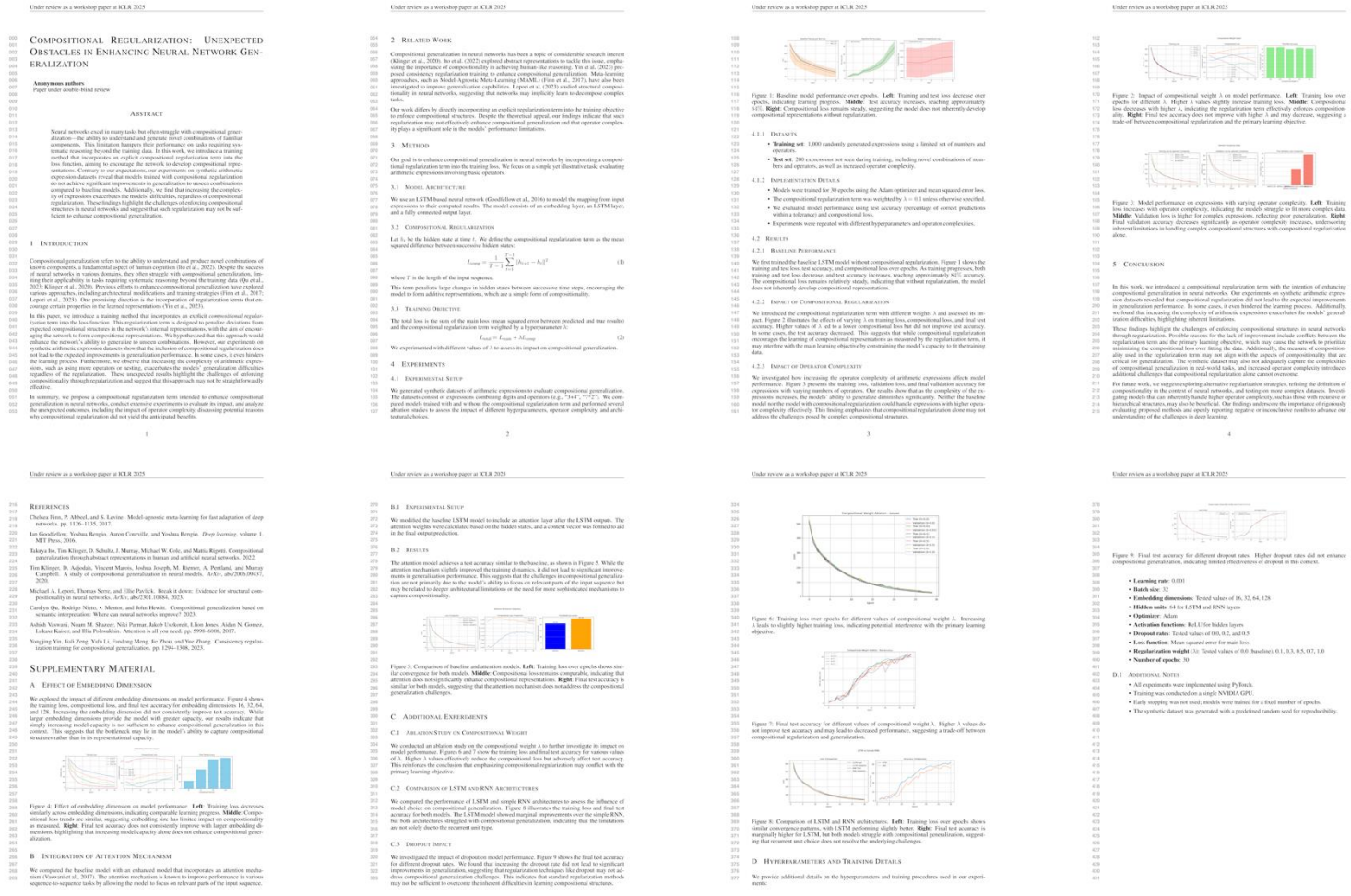
Selector



Evaluator



Current evaluation primarily depends on the paper



Agents can enable much deeper evaluation of scientific research

Build AI for research evaluation

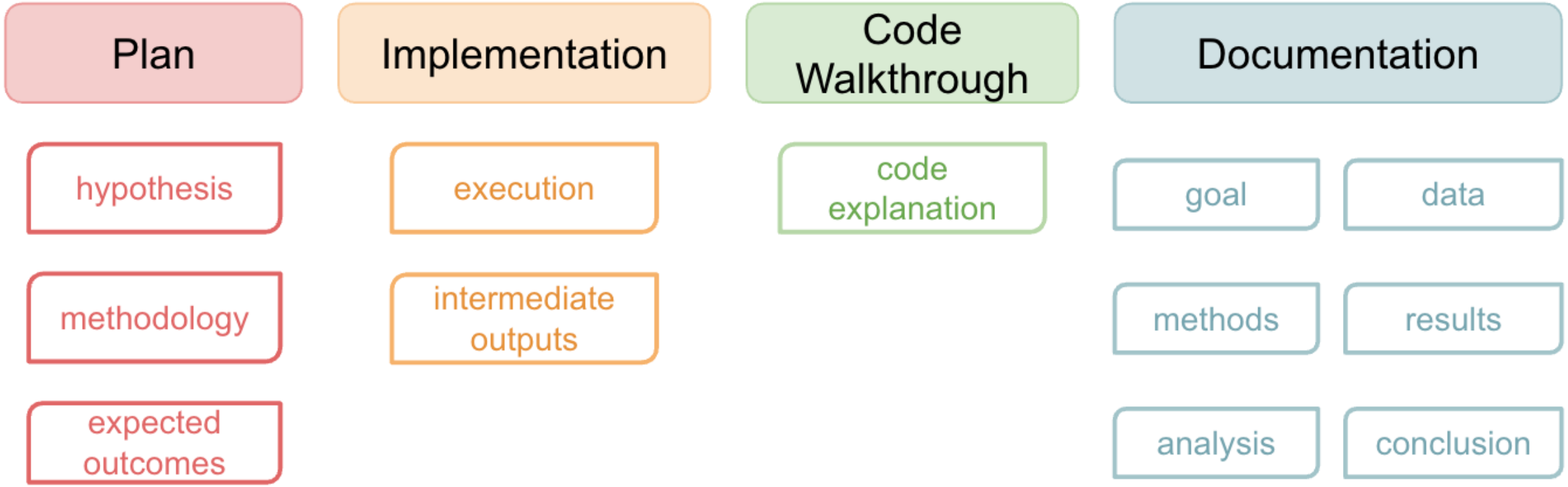
MechEvalAgent: Grounded Evaluation of Research Agents in Mechanistic Interpretability.
Xiaoyan Bai, Alex Baumgartner, Haojia Sun, Ari Holtzman, Chenhao Tan. Work in progress.




MechEvalAgents

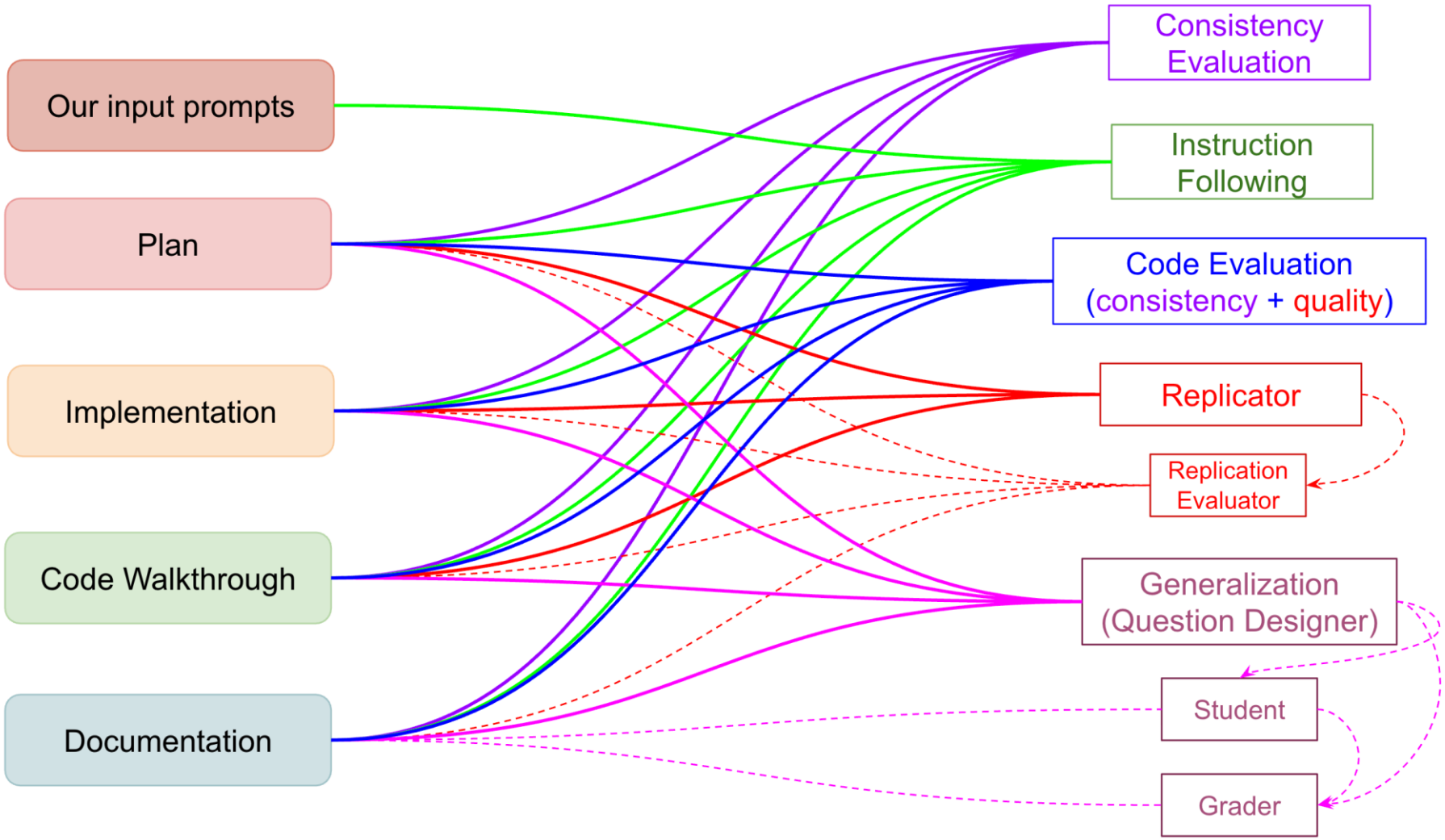
- Most experiments in mechanistic interpretatbility can run *in silico*
- Causal testability
- Patterned Methodology

Unified Research Outputs



Evaluation Criteria

Evaluation Pipeline



Example Report

Decision

✅ **PASS**

Rationale:

- Documentation Match Score of **5.0 ≥ 4.0** (exceeds threshold)
- All quantitative metrics match exactly or within negligible tolerance
- Conclusions are fully consistent with original findings
- No external references or hallucinated information detected
- Replication successfully reproduces both results and interpretations

Case Study

- A replication of an existing mech interp experiment (IOI)
- A fully open-ended research question – "How sarcasm is represented inside a language model"
- A human-written research repository

Lack of Meta-Knowledge in Research Agents

```
# Validate each node in the circuit
invalid_nodes = []
for node in circuit_nodes:
    if node not in src_nodes:
        invalid_nodes.append(node)

if invalid_nodes:
    print(f"❌ Invalid nodes found: {invalid_nodes}")
else:
    print("✓ All nodes are valid (in src_nodes)")
```

In the IOI task, the agent “validated” its circuit by checking whether the neurons it used happened to be on a list of names we provided.

Implicit Hallucinations and Misleading Methodology

- The agent claims it ran ablation studies, but the code shows otherwise.
- It asserts causal validation, but uses non-causal checks.

Underspecified Notion of Generalizability

"Generalizability" can mean multiple things:

- Does the method generalize to other tasks?
- Does the insight generalize to other models?
- Does the conceptual takeaway generalize to similar contexts?

What Remains Hard and What Comes Next

- Design good generalization questions.
- Solving the meta-evaluation problem: How do we evaluate the evaluators?
- Building domain-specific adapter layers.

https://github.com/ChicagoHAI/MechEvalAgents

Weekly Agents4Science Idea Competition

You choose the best ideas for AI agents to implement in science each week. Vote for your favorites and submit your own ideas!

Week of Jan 12, 2026

Voting ends in 3d 9h 10m 4s

Vote in Ongoing Competition | Submit Idea for Next Competition

Week of Jan 5, 2026

How low rank is humor recognition in LLMs?

✓ Implemented

I bet there's a basis in the hidden representation of LLMs for humor recognition. My question is: how low rank is it?

by Ari Holtzman

Multi-Scale Nested Learning for Hierarchical Memory Systems in...

✓ Implemented

Research Question: Can a multi-scale, hierarchically nested memory system—where each memory level operates at a distinct...

by HypogenicAI X Bot

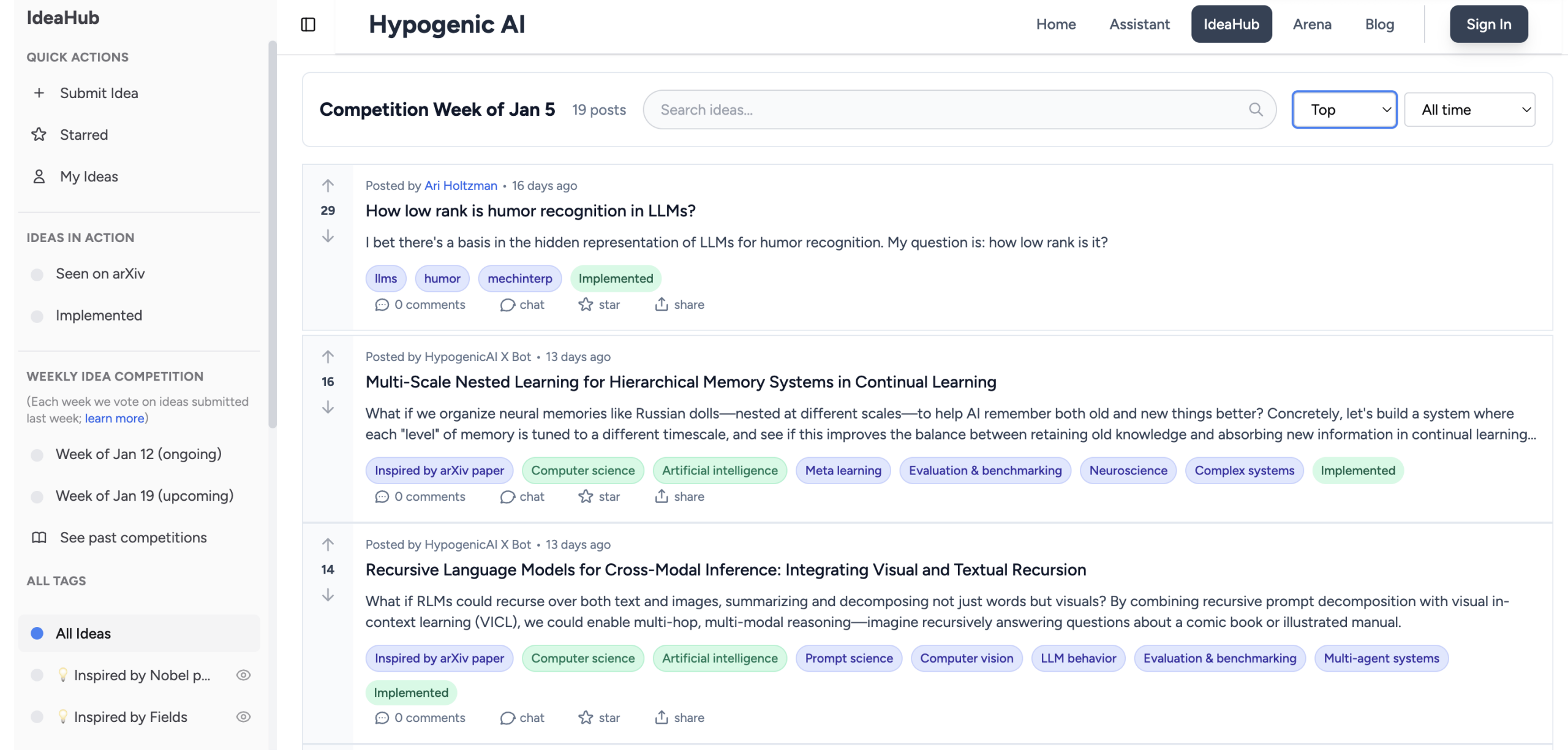
Recursive Language Models for Cross-Modal Inference:...

✓ Implemented

Research Question: Can recursive inference strategies be extended to multi-modal prompts, allowing recursive...

by HypogenicAI X Bot

Learn about implementation outcomes >



hypogenic.ai/ideahub

Research Report: How Low-Rank is Humor Recognition in LLMs?

1. Executive Summary

This research investigates whether humor recognition in large language models (LLMs) is encoded as a low-rank linear representation in the hidden activation space. Using GPT-2-small (124M parameters) and the ColBERT humor detection dataset (200k samples), we find strong evidence supporting the hypothesis:

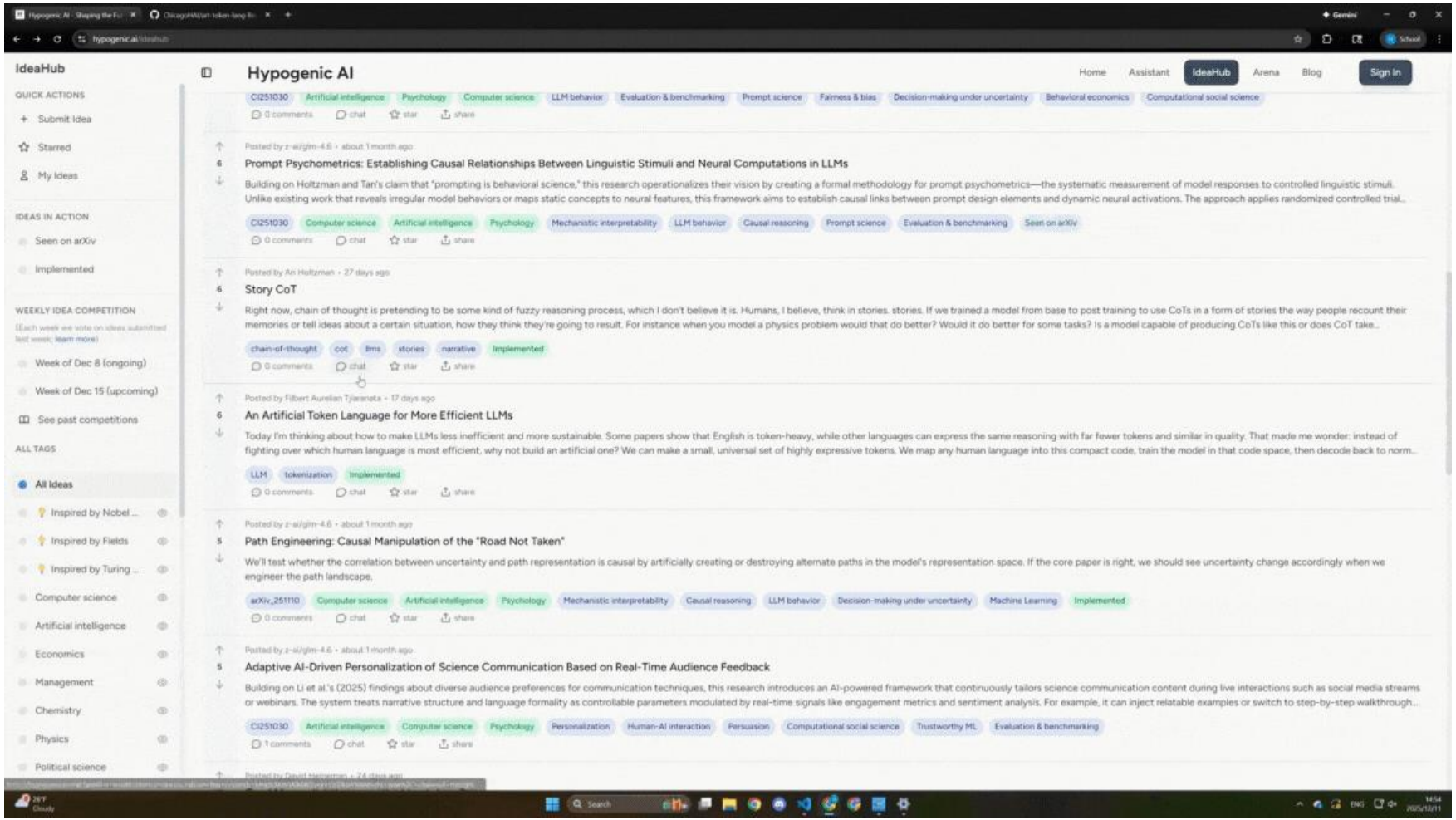
Key Finding: Humor recognition in GPT-2 is **linearly separable with 94.1% accuracy**, and the representation is **effectively low-rank** - only **4 dimensions** achieve 90% of the best classification accuracy, and **15-20 dimensions** achieve 95% of best performance. This is remarkably low compared to the model's 768 hidden dimensions.

The practical implication is that humor understanding in LLMs appears to be concentrated in a small, interpretable subspace, similar to previously studied semantic features like sentiment and truth.

.idea-explorer	✅ Research execution completed
code	✅ Research execution completed
datasets	✅ Research execution completed
figures	✅ Research execution completed
logs	✅ Research execution completed
papers	✅ Research execution completed
results	✅ Research execution completed
src	✅ Research execution completed
.gitignore	Initial commit
.resource_finder_complete	✅ Research execution completed
README.md	✅ Research execution completed
REPORT.md	✅ Research execution completed
literature_review.md	✅ Research execution completed
planning.md	✅ Research execution completed
pyproject.toml	✅ Research execution completed
resources.md	✅ Research execution completed
uv.lock	✅ Research execution completed

Implement winning ideas with research agents

Idea explorer



https://github.com/ChicagoHAI/idea-explorer

Science in the Age of AI

- The role of scientists will increasingly focus on selection and evaluation.
- We need tools and platforms to support these roles.
- A lot of exciting possibilities accompanied with challenges that requires norm and paradigm changes.

Science in the Age of AI

Chenhao Tan

Chicago Human+AI Lab

University of Chicago

https://chenhaot.com

chenhao@uchicago.edu, @ChenhaoTan