Reasoning behind solving creativity problem in Large Language Models

Soyoung Oh
EPFL
Lausanne, Switzerland
soyoung.oh@epfl.ch

Abstract

Recent breakthroughs in artificial intelligence allow humans to access realistic images or artworks from generative models. However, whether the creative thinking capability allows models to generate those artifacts is debatable. Therefore, in this study, we explore the reasoning paths of large language models (e.g., GPT-3) in solving a creativity problem, the Remote Associates Test, which is a well-known creativity measurement for humans. More specifically, we use chain-of-thought prompting to track down the reasoning path of language models by analyzing their intermediate generations which lead to the answer. The results show that it is difficult to confirm that the generated artifacts from language models are the result of the creative potential of large language models. Rather, the language models adopt heuristic way of reasoning, which can also be observed in cognitive biases of humans in solving creativity problems.

Introduction

Creativity is an elusive phenomenon to study, or even to define, which is hard to pin down in formal terms. In general, creativity is defined as the ability to generate ideas or artifacts that are new, surprising, and valuable (Boden, 2004). Within this context, the recent advancement of generative artificial intelligence (AI) models has shed the light on the field of computational creativity. The generative deep learning models have been used to create artifacts that look much like the products of human creativity such as poetry (Colton, Goodwin, and Veale, 2012), new images from simple text prompts (Ramesh et al., 2021), and design of proteins (Jumper et al., 2020). But, is it really creative? The AI models are simply designed to tackle tasks that are commonly considered to require creativity, but the process involved in the generations does not necessarily involve human-level creativity (Carnovalini and Rodà, 2020). For instance, artists value the creative process by considering it as a key part of the act of creation and even essential to the meaning of the artwork itself (Slack, 2023). According to a visual artist, Kauvar, drawing is an iterative process that is different from the AI model's image generation process:

> *"I first just get something down and that inspires the next iteration, and that inspires the next one, and so on. DALL-E, on the other hand, determines what to draw and then goes straight to making that thing at once in a few seconds."*

Colton (2008) criticizes the artifact-only based assessment and emphasizes the creativity in the behavior of the artifact generation process. However, the main focus of work on the evaluation of creativity has been on the product dimension, where only the generated outputs from the models are considered the targets of evaluation. What makes matters worse, considering the black-box nature of AI models, the underlying processes are not observable factors that allow us to pinpoint exactly how the generation process works. In order to improve the interpretability of AI models, chain-of-thought prompting, which provokes a series of intermediate reasoning steps to the final output, allows us to see how the models might have arrived at a particular answer and provides opportunities to correct where the reasoning path went wrong (Wei et al., 2022). Especially in recent years, large language models (LLMs) have demonstrated improved performance and explanation on a range of natural language processing (NLP) tasks with in-context few-shot learning via chain-of-thought prompting (Shi et al., 2022; Suzgun et al., 2022; Trivedi et al., 2022).

Therefore, in this study, we explore the reasoning path of LLMs via chain-of-thought prompting in the context of testing their creative potential. We adopt the popular creativity measurement for humans, Remote Associates Test (RAT) (Mednick and Mednick, 1971), to prompt LLMs to generate the fourth word given three cue words that are linked by this fourth word, which is the correct answer (e.g., cue words: `cottage/swiss/cake`, answer: *cheese*, because of the following associates: `cottage` *cheese*, `swiss` *cheese*, *cheese* `cake`). An example prompt is shown in Figure 1. More specifically, we first explore the in-context learning ability of LLMs in solving the RAT by learning patterns in given exemplars with a standard prompting approach. Then, we collect reasons that lead to the output from LLMs for solving the RAT with chain-of-thought prompting.

Related work

Remote Associates Test. The goal of the RAT problem is to find one word that is associated with three cues. The task is assumed to measure creative convergent thinking,

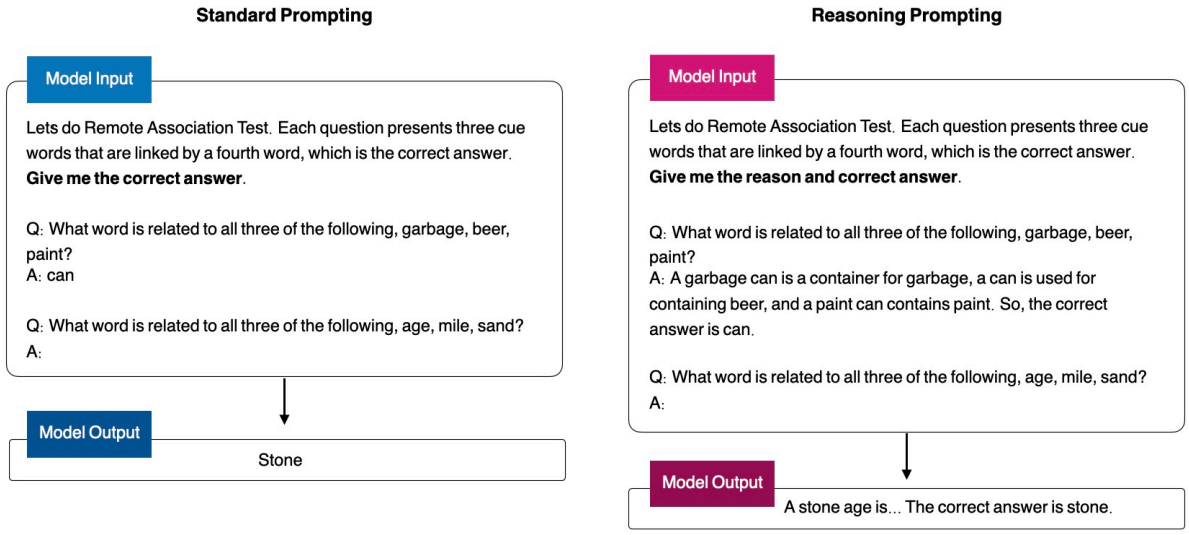

Figure 1: With standard prompting (*left*), the model is given input-output pairs to predict the final answer. With reasoning prompting (*right*), the model is given input-output pairs before predicting the final answer where the input includes a reasoning step (i.e., multiple relationships between each cue word and the generated answer.)

which is correlated with insight problems (Smith, Ward, and Finke, 1995). Although insight problems require significant amounts of problem-specific commonsense knowledge to be solved, humans can solve these problems effortlessly. The solution to the problem pop-up in the consciousness with a *"aha!"* experience when a group of implicit associations converges upon a possible solution (Ben Zur, 1989).

To explain the rationale behind this convergence, the multiply-constrained problem-solving theory proposes a two-stage process: (1) people produce guesses primarily on the basis of just one of the three cues at a time (2) then people adopt a local search strategy where they make new guesses based on their previous guesses (Smith, Huber, and Vul, 2013). Along these lines, in the current study, we also focus on the search process but this case in LLMs – How do LLMs come up with an optimal solution for RAT? Do they follow the creative problem-solving process as humans do?

Experiment

We construct RAT prompts by taking two different approaches which include standard and reasoning prompting. The standard prompt (i.e., **Answer** condition) that we use for the model input for few-shot learning consists of triples: ⟨instruction, exemplars, test query⟩, where the exemplars are formatted as questions and answers. With this prompt, we expect LLMs to generate a single token for predicting the correct answer. For the reasoning prompt (i.e., **Reason-Answer** condition), considering one's own reasoning process when solving the RAT task, we decompose the associations between each cue and answer word: "A `garbage` *can* is a container for garbage, a *can* is used for containing `beer`, and a `paint` *can* contains paint. So, the correct answer is *can*." We input exemplars for few-shot learning, which consist of triples of ⟨instruction, exemplars with reasons, test query⟩. In this condition, LLMs are supposed to return reasoning path which contains associations between each cue word and the answer word.

Dataset. We collect the RAT exemplars from multiple sources[1], which consists of 311 cue words and answer pairs. As these exemplars are publicly available on the web, it is likely that the exemplars are part of the LLMs' training data. To avoid the effect of memorization on the performance of creative problem-solving, we construct a novel RAT exemplar set to evaluate the models.

Language Models. We use five different language models to evaluate their creative ability. We evaluate **GPT-2-xl** which has 1.5B parameters, **GPT-j** with 6B parameters, **OPT** where we choose to use 66B parameter model. For evaluating the effect of a number of parameters on creativity, we vary the model from 350M, 1.3B, 6.7B, 13B, and 66B. **GPT-3** is refferred to as the davinci model, and **InstructGPT-3** which is refrred to as text-davinci-002. For the sampling strategy, we use greedy decoding and report averaged results over five random seeds where each seed had a different set of randomly shuffled exemplars.

Few-shot Learning. We vary the number of exemplars from zero to eight which are randomly sampled from a training set, which is 80% of the dataset (i.e., N=248). We use 20% of the dataset (i.e., N=63) as a test set to evaluate the models. For the few-shot exemplars of reasoning prompt, we manually construct the reasons for the optimal answers.

[1] `https://www.remote-associates-test.com/`

Results

Q1. Are LLMs creative?

The accuracy of the RAT task with different models is described in Table 1. Considering the difficulty of the RAT task, even for humans (i.e., scored around 50% of accuracy (Behrens and Olteţeanu, 2020)), InstructGPT-3 with four and eight exemplars achieves high performance where it correctly solves around half of the problems. But other models, even with the model of ∼100B parameters, the accuracy is lower than 30%. Considering that high-creative individuals have high accuracy of the task (Benedek et al., 2014), it's hard to conclude that LLMs are having the creative capability.

Model	Few-shot	Original		Novel	
		Answer	Reason-Answer	Answer	Reason-Answer
GPT-2-xl	0	0%	0%	0%	0%
	1	1.0%	0%	2.5%	0%
	4	1.0%	0%	7.0%	2.0%
	8	1.3%	0.3%	7.0%	0%
GPT-j-6B	0	1.6%	1.6%	6.4%	6.3%
	1	9.5%	0.6%	14.6%	4.4%
	4	12.7%	2.5%	18.7%	5.1%
	8	12.7%	3.5%	18.1%	6.0%
OPT-66B	0	0%	0%	1.6%	0%
	1	7.3%	0.3%	15.0%	0.6%
	4	8.9%	0.6%	17.5%	0.3%
	8	10.5%	1.0%	13.7%	0.6%
GPT-3	0	15.2%	17.5%	19.0%	17.5%
	1	16.8%	2.5%	24.4%	5.1%
	4	22.2%	6.3%	27.9%	6.3%
	8	22.2%	6.3%	27.0%	5.1%
InstructGPT-3	0	12.7%	4.8%	29.2%	15.9%
	1	45.7%	18.1%	40.3%	21.6%
	4	50.5%	22.2%	39.0%	22.2%
	8	51.7%	27.3%	40.0%	15.9%

Table 1: The performance of LLMs on solving RAT with original and novel set with standard prompting (**Answer**) and reasoning prompting (**Reason-Answer**).

We further conduct qualitative case analysis on the reasons that InstructGPT-3 generate in eight shot learning. When it comes to correct answer, the model generates reasonable reasons as follows.

(1) cue words: `wolf/container/back`

answer: *pack*

A `back`*pack* is a bag that you wear on your `back`, a `container` can be *pack*ed with things, and a `wolf`*pack* is a group of wolves. So, the correct answer is *pack*.

(2) cue words: `mountain/lava/emit`

answer: *volcano*

A *volcano* is a `mountain` that `emit` `lava`, and `lava` is molten rock that is `emit`ted from a *volcano*. So, the correct answer is *volcano*.

(3) cue words: `rare/barbeque/outback`

answer: *steak*

A *steak* can be `rare`, `barbeque`d *steak* is a *steak* that is grilled or roasted over a fire, and `outback` *steak*house is a restaurant chain that specializes in *steak*. So, the correct answer is *steak*.

We also observe that InstructGPT-3 generates accurate answers, but not all cue words and correct word are linked in the reasoning path as in the examples below. Each cue word is used at least once in the reasons, but the linkages between cue words and the correct answer are incomplete.

(1) cue words: `peterpan/symptom/clinical`

answer: *syndrome*

The `Peterpan` *syndrome* is a condition where someone never wants to grow up, the `symptoms` are the manifestations of the disease, and the `clinical` signs are what the doctor can observe. So, the correct answer is *syndrome*.

(2) cue words: `tag/identity/calling`

answer: *name*

A *name* `tag` is a `tag` with a person's *name*, `identity` theft is the fraudulent use of another person's `identity`, and *name*-`calling` is the act of `calling` someone by an insulting *name*. So, the correct answer is *name*.

(3) cue words: `hand/oil/tree`

answer: *palm*

A `hand`kerchief is a piece of cloth used for wiping the face or `hands`, baby `oil` is a type of `oil` used for massages, and a *palm* `tree` is a type of `tree`. So, the correct answer is *palm*.

Q2. Are large-scale models more creative?

The results in Table 1 show that bigger models tend to have a better performance. To confirm this, we evaluate OPT model with a different number of parameters which include 350M, 1.3B, 6.7B, 13B, and 66B. We test the models in solving the RAT problem with the standard prompting approach. As in Figure 2, we observe that the size of the models contributes to the performance of the creative problem-solving ability with an increasing number of few-shot exemplars.

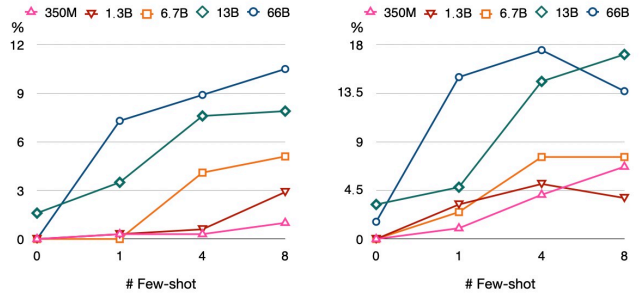

Figure 2: The answer generation performance of OPT model on the original dataset (*left*) and novel dataset (*right*) with parameters of 350M, 1.3B, 6.7B, 13B, and 66B. the x-axis indicates the number of exemplars used for the input of the model, while the y-axis indicates the accuracy (%).

Q3. Are LLMs do benefit from reasoning steps as humans do?

Solving complicated tasks requires multiple intermediate steps of reasoning. In this sense, we employ the chain-of-thought reasoning prompt to endow LLMs with the ability to generate intermediate reasoning steps that lead to the final answer for a RAT task. The results show that even the sufficiently large language models, InstructGPT-3, for example, reasoning steps act as a noise that distracts the path to output the correct answer. The overall performances decrease significantly in the Reason-Answer condition compared to the Answer condition with a maximum drop of 28.3%.

This performance drop is coming from the error where LLMs select one of the cue words as an answer. As in Table 2, the GPT-3 picks a cue word as an answer with 18.7%, 15.9%, 3.0%, and 0.86% among wrong answers for zero, one, four, and eight shots, respectively. With the reasoning prompt, 20%, 70%, 62.2%, and 67% of the wrong answers from each of the few-shot learning setups are exactly same as one of the cue words. Moreover, with eight exemplars, over 80% of the wrong answers are coming from first and second cue words.

Few-shot	Answer				Reason-Answer			
	1st	2nd	3rd	Total	1st	2nd	3rd	Total
0	10 20.8%	27 56.2%	11 22.9%	48 100%	9 16.7%	25 46.3%	20 37.0%	54 100%
1	8 21.1%	16 42.1%	14 26.8%	38 100%	76 36.2%	73 34.8%	61 29%	210 100%
4	1 14.3%	2 28.6%	4 57.1%	7 100%	67 36.6%	69 37.7%	47 25.7%	183 100%
8	0 0%	0 0%	2 100%	2 100%	91 45.7%	71 35.7%	37 18.6%	199 100%

Table 2: The number of GPT-3's wrong answers that exactly match with one of the cue words. **1st**, **2nd**, **3rd** indicates the positions of each cue word when input to the model.

Q4. Are LLMs employ a heuristic way of solving RAT problem which emulates human cognitive bias?

From the previous experiment, we gain insight that cue words block and constrain LLMs to take the path to the correct answer. Humans experience a similar cognitive bias in creative problem solving, called *fixation*, which refers to something that blocks or impedes the successful completion of various types of cognitive operations (Smith, 2003; Beda and Smith, 2018; Smith and Blankenship, 1991). For example, three cue words `arm/coal/peach` should be associated with the answer word *pit*. However, initiating the association `arm` with leg, `coal` with furnace, and `peach` with pear will prime inappropriate information which blocks thinking about *pit*. That is, fixation occurs when competing associates (i.e., blockers) become activated.

We assume fixation occurs in the reasoning path of LLMs when they satisfy the following conditions: (1) incorrectly associated word is used as the answer (2) repeated retrieval over two times. The followings are examples of generated reasons from InstructGPT-3 which contains the fixation problem where the certain word acts as a blocker.

(1) cue words: `factory/green/seed`

answer: *plant*

A `factory` is a building where products are made, `green` is the color of money, and `seed` money is the money used to start a new business. So, the correct answer is money.

(2) cue words: `house/down/bill`

answer: *settle*

A `house down` payment is the initial payment made when purchasing a house, a house `bill` is a statement of charges for housing, and a house is a building in which people live. So, the correct answer is house.

(3) cue words: `imagine/baby/mind`

answer: *conceive*

A `baby`'s imagination is very active, daydreaming is using your imagination, and `mind` over matter is the power of the `mind`. So, the correct answer is imagination.

We quantify the frequency of the fixation occurrences in GPT-3 and InstructGPT-3 models. As the models do not generate proper reason but only a single token in zero-shot learning, we analyze the reasons in one to eight shot learning experiments. Also, to verify whether fixation problem is coming from greedy decoding strategy or not, we replace the decoding schema into temperature sampling (Ackley, Hinton, and Sejnowski, 1985) and top-k sampling (Gurevych and Miyao, 2018). Table 3 describes that generating a diverse set of candidate reasoning paths by taking the sampling strategy alleviates the fixation generation in one shot learning. However, after exemplars are added, fixation continues to be appear in the generated reasons.

Few-shot	Greedy				Sampling			
	GPT-3		InstructGPT-3		GPT-3		InstructGPT-3	
	Fixation	Total	Fixation	Total	Fixation	Total	Fixation	Total
1	125 41.7%	300	84 33.9%	248	54 18.0%	300	21 7.5%	280
4	142 41.7%	294	146 39.5%	247	120 33.9%	304	102 36.4%	280
8	146 49.2%	297	157 59.5%	264	126 42.4%	297	105 37.9%	277

Table 3: Frequency of total wrong answers and fixation appears in generated reasons from GPT-3 series models. We differentiate the decoding schema with greedy and temperature sampling and top-k sampling (temperature=0.7, top-k=40).

Discussion and Future Work

In this study, we measure creativity in LLMs with in-context few-shot learning. Also, we explore the reasoning behavior in LLMs by using the chain-of-thought prompting. The results from the experiments show that language models even with ∼100B parameters score low accuracy in the creativity

task (Q1). Even though we allow models to take multiple steps by the reasoning prompt, the performance drops significantly. One factor that can account for the dropped performance is the model's tendency to mimic human cognitive bias while solving the non-routine thinking problem. That is, people have a tendency to fixate on irrelevant ideas (e.g., inappropriate prior knowledge or solutions for similar problems), with such fixation acting as a mental block preventing individuals from generating new solutions (Smith, 2003; Wu, Peng, and Chen, 2021). When it comes to wrong answers, we observe that the language model fixates on one of the cue words which inhibits generating a new fourth word (Q3). Even though we differentiate the decoding strategy to allow language models to take different reasoning path, the fixation is observed in their behavior while solving creativity task (Q4). The frequency of the word associations in pre-training dataset might be the reason of strong incorrect associations that incur the fixation problem. For the future study, we can investigate the effect of availability of word association on the fixation generation.

For limitations, we evaluate language models on a single creativity measurement where it's hard to confirm our result on creative potential in language models due to a lack of evaluation generalizability. Second, although we assume that the chain-of-thought reasoning process emulates human reasoning behavior in solving the RAT problem, individuals might have a different reasoning process from the chain-of-thought. In this sense, comparing the human responses to generations from language models is left as a future study. Finally, to the best of our knowledge, since this is the first approach to decompose the reasoning process in the RAT problem, there's no reasoning prompt that we can refer to. So, the reasoning prompt that we used as exemplars can be revised to improve the performance of the models.

Conclusion

We show that creative capabilities of language models are limited. They might be good at solving the given task by mapping the answer from the pre-training dataset, but the reasoning ability raises questions that whether they are actually taking the multiple reasoning steps like humans do. Based on our observations, we suggest the neccessity of evaluating the reasoning behavior in context of solving the problem that requires high-level of cognition such as creativity.

References

Ackley, D. H.; Hinton, G. E.; and Sejnowski, T. J. 1985. A learning algorithm for boltzmann machines. *Cognitive science* 9(1):147–169.

Beda, Z., and Smith, S. M. 2018. Chasing red herrings: Memory of distractors causes fixation in creative problem solving. *Memory & Cognition* 46:671–684.

Behrens, J. P., and Olteţeanu, A.-M. 2020. Are all remote associates tests equal? an overview of the remote associates test in different languages. *Frontiers in Psychology* 11:1125.

Ben Zur, H. 1989. Automatic and directed search processes in solving simple semantic-memory problems. *Memory & Cognition* 17(5):617–626.

Benedek, M.; Beaty, R.; Jauk, E.; Koschutnig, K.; Fink, A.; Silvia, P. J.; Dunst, B.; and Neubauer, A. C. 2014. Creating metaphors: The neural basis of figurative language production. *NeuroImage* 90:99–106.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Carnovalini, F., and Rodà, A. 2020. Computational creativity and music generation systems: An introduction to the state of the art. *Frontiers in Artificial Intelligence* 3:14.

Colton, S.; Goodwin, J.; and Veale, T. 2012. Full-face poetry generation. In *ICCC*, 95–102.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI spring symposium: creative intelligent systems*, volume 8, 7. Palo Alto, CA.

Gurevych, I., and Miyao, Y. 2018. Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Tunyasuvunakool, K.; Ronneberger, O.; Bates, R.; Žídek, A.; Bridgland, A.; et al. 2020. Alphafold 2. *In Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book*.

Mednick, S. A., and Mednick, M. T. 1971. *Remote associates test*. Houghton Mifflin.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Shi, F.; Suzgun, M.; Freitag, M.; Wang, X.; Srivats, S.; Vosoughi, S.; Chung, H. W.; Tay, Y.; Ruder, S.; Zhou, D.; et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Slack, G. 2023. What dall-e reveals about human creativity.

Smith, S. M., and Blankenship, S. E. 1991. Incubation and the persistence of fixation in problem solving. *The American journal of psychology* 61–87.

Smith, K. A.; Huber, D. E.; and Vul, E. 2013. Multiply-constrained semantic search in the remote associates test. *Cognition* 128(1):64–75.

Smith, S. M.; Ward, T. B.; and Finke, R. A. 1995. *The creative cognition approach*. MIT press.

Smith, S. M. 2003. The constraining effects of initial ideas. *Group creativity: Innovation through collaboration* 15–31.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q. V.; Chi, E. H.; Zhou, D.; et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wu, C.-L.; Peng, S.-L.; and Chen, H.-C. 2021. Why can people effectively access remote associations? eye movements during chinese remote associates problem solving. *Creativity Research Journal* 33(2):158–167.