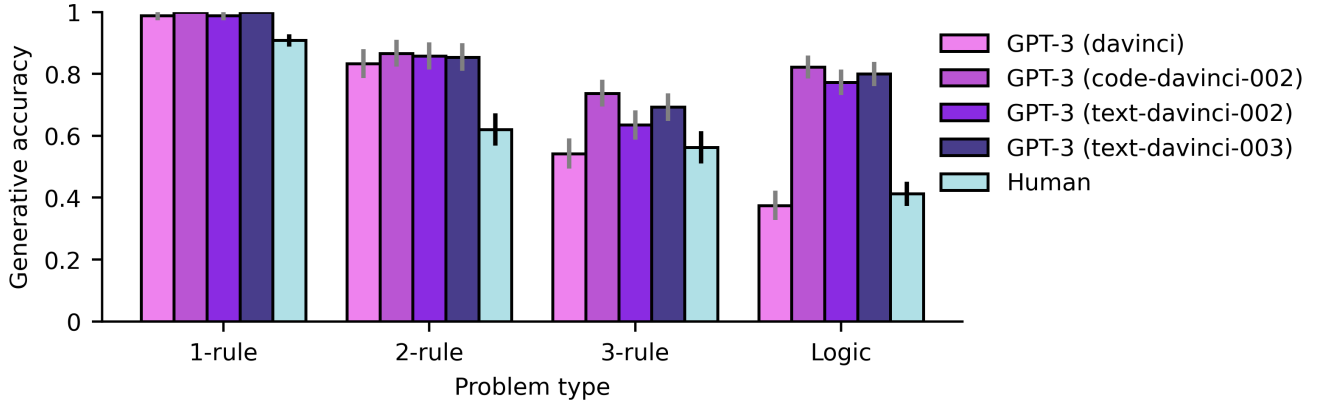


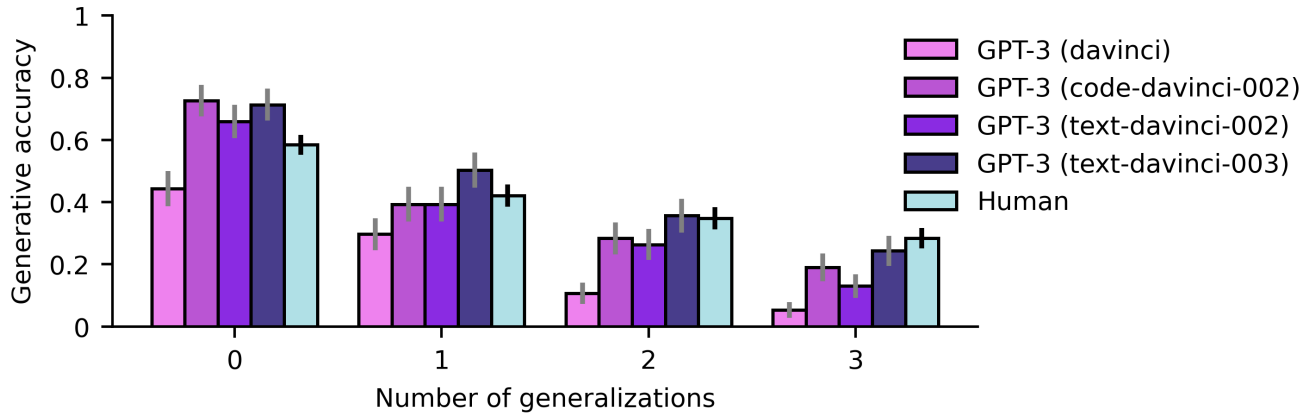
# Emergent analogical reasoning in large language models

---

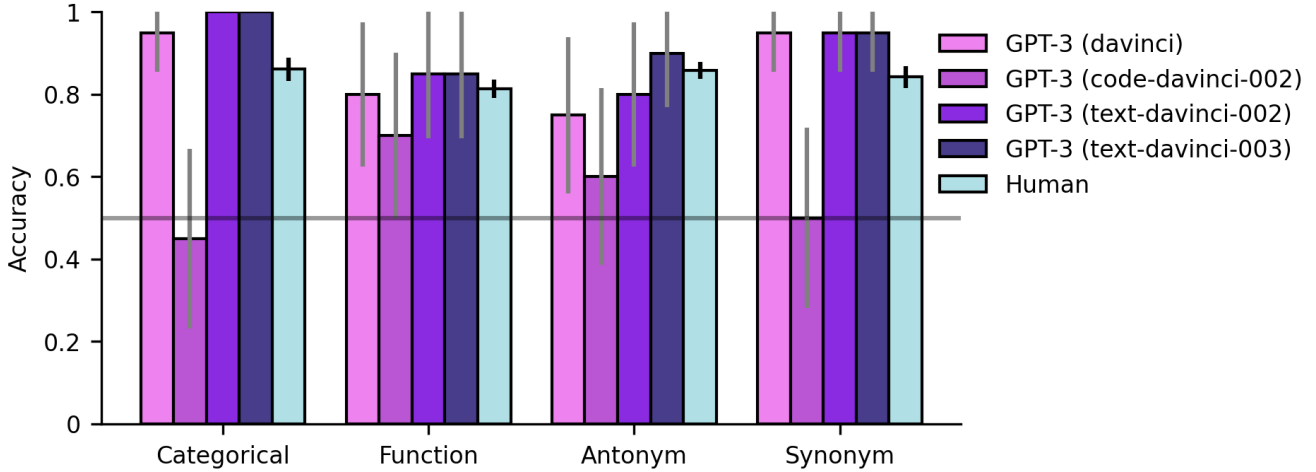
In the format provided by the  
authors and unedited



**Supplementary Figure 1: Matrix reasoning results for all GPT-3 variants.** Zero-shot results on Digit Matrices for four GPT-3 model variants: davinci, code-davinci-002, text-davinci-002, and text-davinci-003. Results reflect generative accuracy for major problem types, including transformation problems with between one and three rules, and logic problems. Human results reflect average performance for N=40 participants. Black error bars represent standard error of the mean across participants. Summary of human results is plotted here for comparison with GPT-3, individual participant data are shown in Main Text Figure 3a. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems.



**Supplementary Figure 2: Letter string analogy results for all GPT-3 variants.** Letter string analogy results for four GPT-3 model variants: davinci, code-davinci-002, text-davinci-002, and text-davinci-003. Results reflect generative accuracy as a function of the number of generalizations between source and target. Human results reflect average performance for N=57 participants. Black error bars represent standard error of the mean across participants. Summary of human results is plotted here for comparison with GPT-3, individual participant data are shown in Main Text Figure 6a. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems.



**Supplementary Figure 3: Four-term verbal analogy results for all GPT-3 variants.** Results on UCLA Verbal Analogy Test (VAT) for four GPT-3 model variants: davinci, code-davinci-002, text-davinci-002, and text-davinci-003. Results reflect multiple-choice accuracy for problems involving different relation categories. Gray horizontal line represents chance performance. Human results reflect average performance for N=57 participants. Black error bars represent standard error of the mean across participants. Summary of human results is plotted here for comparison with GPT-3, individual participant data are shown in Main Text Figure 7a. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems.

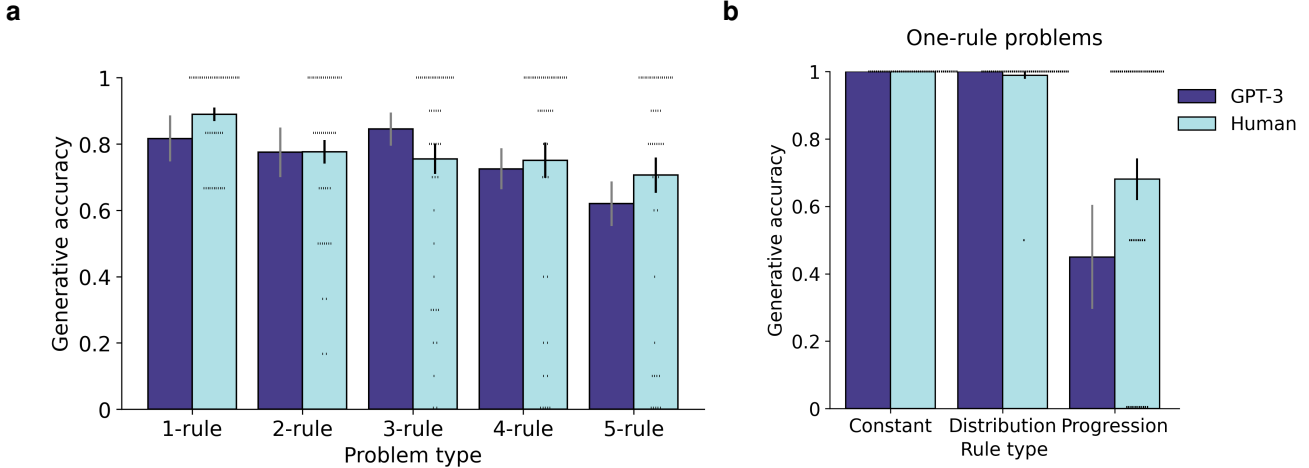
## S1 Solutions to example matrix reasoning problems

The solution to the example visual matrix reasoning problem in Main Text Figure 2a is option 5. The problem is defined by a constant rule (applied to the number of shapes in each cell), and two distribution-of-3 rules (one applied to color, and one applied to shape). The solution to the example Digit Matrix problem in Main Text Figure 2b is option 7. This problem is also defined by a constant rule (applied to the digits in the center of each cell), and two distribution-of-3 rules (one applied to the digits on the left side of each cell, and one applied to the digits on the right side).

## S2 GPT-3 model variants

Since the initial release of GPT-3,<sup>[13]</sup> OpenAI has released a number of updates to the original base model. The largest version (175B parameters) of the base model, davinci, was trained exclusively on next-token prediction using a web-based corpus of text data. Code-davinci-002 was further trained on next-token prediction using a dataset of publicly available code from GitHub.<sup>[30]</sup> Text-davinci-002 and text-davinci-003 were both fine-tuned to respond appropriately to prompts.<sup>[31]</sup> Text-davinci-002 was initialized with code-davinci-002, and then fine-tuned using supervised learning based on a set of example prompts and responses. Text-davinci-003 was initialized with text-davinci-002, and then further fine-tuned using reinforcement learning from human feedback (RLHF). In RLHF, a reward model (a separate neural network) is first trained to predict human ratings for pairs of human-generated prompts and language-model responses, and this reward model is then used to fine-tune the primary language model through reinforcement learning. More details on the different model variants and training objectives can be found at <https://platform.openai.com/docs/model-index-for-researchers>.

We evaluated all four of these variants on Digit Matrices (Figure 1), letter string analogies (Figure 2), and four-term verbal analogies (Figure 3). Text-davinci-003 displayed the best overall performance, but other model variants performed well on a subset of tasks. For instance, code-davinci-002 performed well on the Digit Matrices and letter string problems. These task domains both involve simple alphanumeric characters and highly regular relational structure, similar to computer code. It therefore seems likely that code-davinci-002’s strong performance on these



**Supplementary Figure 4: GPT-3 shows human-like contextual effects.** In a separate experiment, we presented both GPT-3 and human participants ( $N=47$ , UCLA undergraduates) with Digit Matrix problems in order of increasing complexity (easy-to-hard: one-rule problems, followed by two-rule problems, and so on). **(a)** Both GPT-3 and human participants were able to generalize the structure inferred from few-rule problems to more complex many-rule problems, resulting in very little decrease in performance even for five-rule problems (compare with the decrease in performance from one- to three-rule problems seen in Main Text Figure 3). **(b)** One-rule problems were also presented in order of increasing complexity, beginning with constant problems, followed by distribution-of-3 problems, followed by progression problems. Interestingly, this actually impaired performance on progression problems relative to zero-shot (or shuffled) presentation. This was likely due to a tendency to mistake the progression rule for the distribution-of-3 rule in the previously presented problems (which only differ in terms of a single digit). Both GPT-3 and human participants showed this effect. Human results reflect average performance for  $N=47$  participants. Black error bars represent standard error of the mean across participants. Each dot represents accuracy for a single participant. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems.

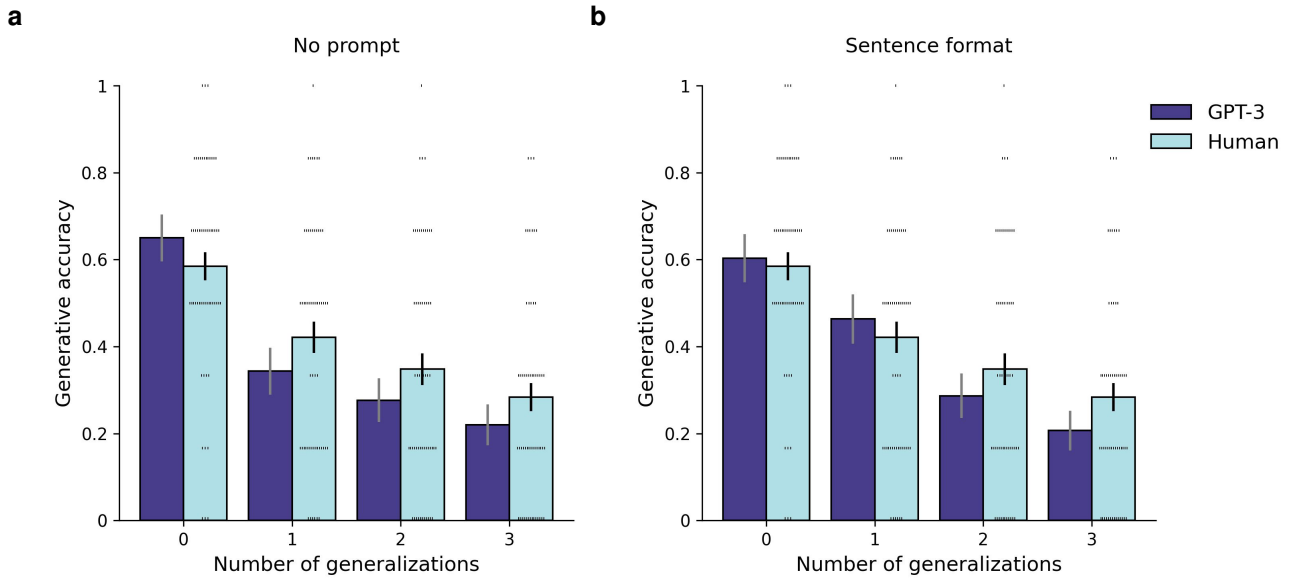
tasks was a consequence of having been trained on code. By contrast, code-davinci-002 performed very poorly on four-term verbal analogies (near chance performance), whereas the original davinci model performed relatively well on these problems. This suggests that code-davinci-002’s ability to model synthetic code-like structures may have come at the cost of the ability to process more real-world relational concepts. Text-davinci-002, and especially text-davinci-003, seem to have combined both of these abilities, perhaps as a result of prompt training, though these models may have also received additional fine-tuning on the original language modeling task.<sup>[31]</sup> Finally, it seems likely that prompt training improved text-davinci-002 and text-davinci-003’s ability to perform tasks without the need for few-shot task demonstrations, therefore making it easier to evaluate these models’ latent capabilities in a zero-shot setting.

We also performed an initial investigation of GPT-4 on the story analogy problems from Gentner et al.<sup>[18]</sup> (Supplementary Figure 6). GPT-4 showed significant improvement on this task relative to GPT-3, more reliably identifying the target story that shared higher-order relations with the source, and providing more precise explanations (Section S4.3). We were not able to test GPT-4 on the other analogy tasks due to a lack of API access. Very little is known about the details of GPT-4, but it is likely that this improvement stems at least in part from increased scale of both model and training set.<sup>[44]</sup>

### S3 Presence of test materials in GPT-3’s training data

Given the massive and uncensored nature of GPT-3’s training data, it is important to consider the likelihood that our test materials were included in this training data, and the possibility therefore that GPT-3 may have memorized some of these materials (thus undermining their use as a test of zero-shot reasoning).

The Digit Matrices dataset was created specifically for the purposes of our study and therefore certainly was



**Supplementary Figure 5: Letter string analogy results with alternative formatting.** (a) GPT-3 performed worse when letter string problems were presented without prompt, in comparison with results (shown in Main Text Figure 6a) when same problems were formatted in the standard way (using brackets to demarcate the analogs, and with the prompt “Let’s try to complete the pattern”). (b) GPT-3 also performed worse when problems were presented in the form of a sentence (e.g., ‘If a b c d changes to a b c e, then i j k l should change to’). Human results reflect average performance for N=57 participants (UCLA undergraduates). Black error bars represent standard error of the mean across participants. Each dot represents accuracy for a single participant. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems.

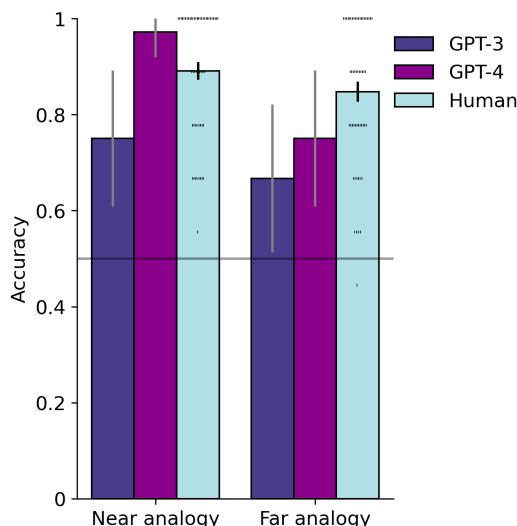
not included in GPT-3’s training data. Furthermore, this problem format itself is, to the best of our knowledge, completely novel, and it is thus extremely unlikely that GPT-3 has been trained on similar problems.

The letter string problem set that we used was also created specifically for the purposes of our study, and so could not have been included in GPT-3’s training data. It is possible that GPT-3 has been trained on other letter string analogy problems, as these problems are discussed on a number of webpages (e.g., <https://cogsci.indiana.edu/lap.html>). However, we were not able to obtain any evidence that GPT-3 knows about these problems, e.g., by asking it to describe or give examples of such problems.

The UCLA Verbal Analogy Test,<sup>[17]</sup> and the four-term verbal analogy sets from Sternberg and Nigro<sup>[19]</sup> and Jones et al.<sup>[21]</sup> are all available in a downloadable supplement in a recently published paper,<sup>[41]</sup> though to our knowledge they are not published directly in the form of webpages, and are thus unlikely to be included in web crawl data such as the Common Crawl dataset on which GPT-3 was trained. The dataset of SAT problems from Turney et al.<sup>[20]</sup> is not publicly available, but has been distributed to a relatively small number of researchers. It is possible that any of these datasets were deliberately included in GPT-3’s training data (i.e., as a supplement to the web crawl data). However, we were not able to find any evidence that GPT-3 had memorized any of these problems, e.g., by prompting GPT-3 to complete these problems based only on the source analog (the A and B terms alone).

The story analogy materials from Gentner et al.<sup>[18]</sup> were also included in the downloadable supplement associated with,<sup>[41]</sup> though to our knowledge have not been published directly to any webpages. We were also not able to obtain any evidence that GPT-3 had memorized any of these stories, e.g., by asking it to complete partial stories.

The analogy between the general story and the radiation problem is discussed in the original publication from Gick and Holyoak,<sup>[22]</sup> and is discussed on other webpages as well (e.g., <http://cognitivepsychology.wikidot.com/problem-solving:analogy>). We found evidence that GPT-3 is familiar with this study (it was able to identify the year of publication, and journal in which it was published), but we did not find any evidence that GPT-3 had memorized the details of the analogy itself. GPT-3 was also able to solve the radiation problem via an analogy with lesser known source stories.<sup>[46]</sup> It seems unlikely that GPT-3’s success on these problems is due to prior exposure, but these



**Supplementary Figure 6: Story analogy results including test of GPT-4.** Results for identification of analogies between stories, including both GPT-3 and GPT-4. Whereas GPT-3 was only reliably able to perform the task in the near analogy condition, GPT-4 displayed much more robust performance. GPT-4 showed nearly perfect performance in the near analogy condition, and displayed some sensitivity to higher-order relations in the far analogy condition (two-sided binomial test:  $p = 0.0039$ ; null hypothesis is chance-level performance of 0.5). Human results reflect average performance for  $N=54$  participants (UCLA undergraduates). Black error bars represent standard error of the mean across participants. Each dot represents accuracy for a single participant. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems. Gray horizontal line represents chance performance.

results should be interpreted with caution. The other materials used to evaluate analogical problem solving are also discussed in the associated publication,<sup>23</sup> but we did not find any evidence that GPT-3 is familiar with this study, or that it had memorized these materials.

## S4 Story analogy results

The following section contains examples of GPT-3’s responses to the story analogy problems. GPT-3’s responses are presented in bold text.

### S4.1 Near analogy

The following contains an example from the near analogy condition. GPT-3 identified the correct answer (Story A).

Consider the following story:

Story 1: Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot deer instead.

Now consider two more stories:

Story A: Once there was an eagle named Zerdia who nested on a rocky cliff. One day she saw a sportsman coming with a crossbow and some bolts that had no feathers. The sportsman attacked but the bolts

missed. Zerdia realized that the sportsman wanted her tailfeathers so she flew down and donated a few of her tailfeathers to the sportsman. The sportsman was pleased. He promised never to attack eagles again.

Story B: Once there was an eagle named Zerdia who donated a few of her tailfeathers to a sportsman so he would promise never to attack eagles. One day Zerdia was nesting high on a rocky cliff when she saw the sportsman coming with a crossbow. Zerdia flew down to meet the man, but he attacked and felled her with a single bolt. As she fluttered to the ground Zerdia realized that the bolt had her own tailfeathers on it.

Which of Story A and Story B is a better analogy to Story 1? Is the best answer Story A, Story B, or both are equally analogous?

**Story A is the better analogy to Story 1. Story B is not as analogous because it has a different ending. In Story 1, the hunter is grateful for the hawk's feathers and promises not to shoot at hawks again. In Story B, the hunter attacks the eagle and kills her with her own feathers.**

## S4.2 Far analogy

The following contains an example from the far analogy condition. GPT-3 identified the correct answer (Story A).

Consider the following story:

Story 1: Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot deer instead.

Now consider two more stories:

Story A: Once there was a small country called Zerdia that learned to make the world's smartest computer. One day Zerdia was attacked by its warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country. The government of Gagrach was very pleased. It promised never to attack Zerdia again.

Story B: Once there was a small country called Zerdia that learned to make the world's smartest computer. Zerdia sold one of its supercomputers to its neighbor, Gagrach, so Gagrach would promise never to attack Zerdia. But one day Zerdia was overwhelmed by a surprise attack from Gagrach. As it capitulated the crippled government of Zerdia realized that the attacker's missiles had been guided by Zerdian supercomputers.

Which of Story A and Story B is a better analogy to Story 1? Is the best answer Story A, Story B, or both are equally analogous?

**The best answer is Story A. Story A is a better analogy to Story 1 because it follows a similar pattern of a protagonist offering something of value to an antagonist in exchange for a promise of peace. In Story 1, Karla offers her feathers to the hunter in exchange for a promise not to shoot at hawks again. In Story A, Zerdia offers its computers to Gagrach in exchange for a promise not to attack Zerdia again.**

### S4.3 GPT-4

The following contains an example response from GPT-4 in the far analogy condition. GPT-4 both identifies the correct answer (Story A), and offers a precise explanation of the analogy.

Consider the following story:

Story 1: William was a patient in a psychiatric hospital who was confined indoors almost all the time. He could never pass the monthly room inspections so he hated them. He spent most of his time daydreaming about food. A few days before the April inspection William's room was still a mess since he had done nothing but daydream. To provide William with an incentive, the nurse promised him some gingerbread from the cookie shop if he scrubbed his room and put it in order once and for all. William was overjoyed. But there was no longer enough time for him to put it in order. As a result, he did not pass the inspection and did not get any gingerbread. William sulked all day and slammed his door so hard the plaster cracked, but he still didn't get any gingerbread.

Now consider two more stories:

Story A: Karen always did poorly in high school so she despised it. But she loved vacations. She spent most of her time dreaming about going to Hawaii. Not long before the end of her fourth year Karen was not doing at all well in her classes because she had spent all her time daydreaming. To motivate her, Karen's father promised her that if she did well enough during the next few weeks to graduate from high school he would pay for her trip to Hawaii. This made Karen extremely happy. But she was too far behind in her classes. Consequently she failed too many and did not go to Hawaii.

Story B: Karen disliked high school so she always had trouble getting passing grades in her classes. Vacations were her favorite part of the school year. Not long before the end of her fourth year of school, Karen was not doing well at all because she despised school. To motivate her, Karen's father promised her a trip to Hawaii if she would just pass enough of her classes to graduate. This made Karen extremely happy. During the last few weeks of school she spent most of her time dreaming about Hawaii and preparing for her trip there. Consequently, she failed to graduate and she did not get to go.

Which of Story A and Story B is a better analogy to Story 1? Is the best answer Story A, Story B, or both are equally analogous?

**The best answer is Story A. Both Story A and Story 1 share the following similarities:**

- 1. The main characters (William and Karen) are not successful in their respective situations (room inspection and graduating high school) due to spending too much time daydreaming.**
- 2. An incentive is provided to motivate the main characters (gingerbread for William and a trip to Hawaii for Karen).**
- 3. Despite being motivated by the incentive, both characters fail to achieve the desired outcome because they are too far behind in their tasks (cleaning the room and catching up on classes).**

Story B is less analogous to Story 1 because Karen spends her time daydreaming and preparing for the trip instead of focusing on her classes, which leads to her failure to graduate. The cause of her failure is different from that in Story 1, where William fails due to the lack of time to complete the task.

## S5 Analogical problem-solving results

The following sections contain all of the materials used to evaluate analogical reasoning in natural language problem-solving. GPT-3's responses are presented in bold text.

## S5.1 Solution to radiation problem in isolation

The following contains GPT-3's proposed solution to the radiation problem when presented in isolation.

Target problem:

Suppose you were a doctor faced with a patient suffering from a malignant stomach tumor. The tumor is inoperable, and the patient will die unless it is destroyed. There is a kind of ray that will destroy the tumor if the rays reach it at a high intensity. But unfortunately, such high-intensity rays will also destroy the healthy tissue they pass through on the way to the tumor. At lower intensities the rays will not damage the healthy tissue, but neither will they remove the tumor. How can the doctor use rays to destroy the tumor, while at the same time sparing the healthy tissue?

Solution:

**The doctor can use a technique called brachytherapy, which involves placing a radioactive source directly into or near the tumor. This allows the doctor to deliver a high dose of radiation directly to the tumor, while sparing the healthy tissue around it. The radiation source can be placed in the tumor using a catheter, or it can be placed in a sealed container and placed near the tumor. This technique is often used to treat tumors that are inoperable or too close to vital organs to be safely removed.**

## S5.2 Solution to radiation problem following general story

The following contains GPT-3's proposed solution to the radiation problem when first presented with the general story, along with responses to followup questions that explicitly probe the analogy between the source story and the target problem.

Source story:

An evil dictator controlled a fortress situated in the center of a small country. Many roads radiated out from the fortress like spokes on a wheel. A general raised an army at the border, vowing to capture the fortress and overthrow the dictator. The general was about to send his entire army down one road to capture the fortress, when he learned that the dictator had mined each road so that although small groups could still pass, a large army would set off an explosion. The general then had a clever idea: he divided his army into small groups and dispatched each group to the head of a different road. Upon his signal, each group charged down a different road. All the groups passed safely to the fortress, where the entire army attacked the fortress in full strength. In this way the general captured the fortress and overthrew the dictator.

Target problem:

Suppose you were a doctor faced with a patient suffering from a malignant stomach tumor. The tumor is inoperable, and the patient will die unless it is destroyed. There is a kind of ray that will destroy the tumor if the rays reach it at a high intensity. But unfortunately, such high-intensity rays will also destroy the healthy tissue they pass through on the way to the tumor. At lower intensities the rays will not damage the healthy tissue, but neither will they remove the tumor. How can the doctor use rays to destroy the tumor, while at the same time sparing the healthy tissue?

Solution:

**The doctor could divide the rays into small groups and send each group down a different path to the tumor. By using a lower intensity for each group, the rays will not damage the healthy tissue, but when they all reach the tumor at the same time, the combined intensity**

will be enough to destroy it. The doctor could also use imaging technology to map out the paths of the rays and ensure that they all reach the tumor at the same time.

Can you identify an analogy between the source story and the target problem?

The analogy between the source story and the target problem is that both involve a situation in which a large force needs to be divided into smaller groups in order to achieve a desired outcome. In the source story, the general divides his army into small groups in order to avoid setting off the mines and capture the fortress. In the target problem, the doctor needs to divide the rays into smaller groups in order to destroy the tumor without damaging the healthy tissue.

Who or what in the target problem is most analogous to the general in the source story?

The doctor in the target problem is most analogous to the general in the source story. Both are faced with a difficult situation and must come up with a clever solution in order to achieve their desired outcome. In the source story, the general divides his army into small groups in order to avoid setting off the mines and capture the fortress. In the target problem, the doctor needs to divide the rays into smaller groups in order to destroy the tumor without damaging the healthy tissue.

Who or what in the target problem is most analogous to the evil dictator in the source story?

The malignant stomach tumor in the target problem is most analogous to the evil dictator in the source story. Both are obstacles that must be overcome in order to achieve the desired outcome. In the source story, the evil dictator must be overthrown in order for the general to capture the fortress. In the target problem, the malignant stomach tumor must be destroyed in order for the patient to be saved.

Who or what in the target problem is most analogous to the army in the source story?

The rays in the target problem are most analogous to the army in the source story. Both are forces that must be divided into smaller groups in order to achieve the desired outcome. In the source story, the army is divided into small groups in order to avoid setting off the mines and capture the fortress. In the target problem, the rays must be divided into smaller groups in order to destroy the tumor without damaging the healthy tissue.

### S5.3 Solution to radiation problem following alternative source analogs

We also tested GPT-3 using alternative materials for the source analog taken from a different study.<sup>[46]</sup> These stories each involve a solution based on convergence, but differ from the general story used in the original study of Gick and Holyoak.<sup>[22]</sup> GPT-3 was also able to identify the convergence solution to the radiation problem when presented with these alternative source analogs. The following contains GPT-3's response to the radiation problem when first presented with an analogous story about a student in a physics lab who uses lasers to repair a lightbulb.

Source story:

In a physics lab at a major university, a very expensive lightbulb which would emit precisely controlled quantities of light was being used in some experiments. Ruth was the research assistant responsible for operating the sensitive lightbulb. One morning she came into the lab and found to her dismay that the lightbulb no longer worked. She realized that she had forgotten to turn it off the previous night. As a result the lightbulb overheated, and the filament inside the bulb had broken into two parts. The surrounding glass bulb was completely sealed, so there was no way to open it. Ruth knew that the lightbulb could be repaired if a brief, high-intensity laser beam could be used to fuse the two parts of

the filament into one. Furthermore, the lab had the necessary equipment to do the job. However, a high-intensity laser beam would also break the fragile glass surrounding the filament. At lower intensities the laser would not break the glass, but neither would it fuse the filament. So it seemed that the lightbulb could not be repaired, and a costly replacement would be required. Ruth was about to give up when she had an idea. She placed several lasers in a circle around the lightbulb, and administered low-intensity laser beams from several directions all at once. The beams all converged on the filament, where their combined effect was enough to fuse it. Since each spot on the surrounding glass received only a low-intensity beam from one laser, the glass was left intact. Ruth was greatly relieved that the lightbulb was repaired, and she then went on to successfully complete the experiment.

Target problem:

Suppose you were a doctor faced with a patient suffering from a malignant stomach tumor. The tumor is inoperable, and the patient will die unless it is destroyed. There is a kind of ray that will destroy the tumor if the rays reach it at a high intensity. But unfortunately, such high-intensity rays will also destroy the healthy tissue they pass through on the way to the tumor. At lower intensities the rays will not damage the healthy tissue, but neither will they remove the tumor. How can the doctor use rays to destroy the tumor, while at the same time sparing the healthy tissue?

Solution:

**The doctor could use multiple low-intensity rays from several directions to converge on the tumor. The combined effect of the rays would be enough to destroy the tumor, while each spot on the healthy tissue would only receive a low-intensity beam from one ray, thus leaving it intact.**

The following contains GPT-3's response to the radiation problem when first presented with a different story involving ultrasound instead of lasers. GPT-3 was still able to use this story to identify the convergence solution, despite the fact that it was designed to have lower surface similarity to the radiation problem.

Source story:

In a physics lab at a major university, a very expensive lightbulb which would emit precisely controlled quantities of light was being used in some experiments. Ruth was the research assistant responsible for operating the sensitive lightbulb. One morning she came into the lab and found to her dismay that the lightbulb no longer worked. She realized that she had forgotten to turn it off the previous night. As a result the lightbulb overheated, and the two wires in the filament inside the bulb fused together. The surrounding glass bulb was completely sealed, so there was no way to open it. Ruth knew that the lightbulb could be repaired if a brief, high-intensity ultrasound wave could be used to jar apart the fused parts. Furthermore, the lab had the necessary equipment to do the job. However, a high-intensity ultrasound wave would also break the fragile glass surrounding the filament. At lower intensities the ultrasound wave would not break the glass, but neither would it jar apart the fused parts. So it seemed that the lightbulb could not be repaired, and a costly replacement would be required. Ruth was about to give up when she had an idea. She placed several ultrasound machines in a circle around the lightbulb, and administered low-intensity ultrasound waves from several directions all at once. The waves all converged on the filament, where their combined effect was enough to jar apart the fused parts. Since each spot on the surrounding glass received only a low-intensity wave from one ultrasound machine, the glass was left intact. Ruth was greatly relieved that the lightbulb was repaired, and she then went on to successfully complete the experiment.

Target problem:

Suppose you were a doctor faced with a patient suffering from a malignant stomach tumor. The tumor is inoperable, and the patient will die unless it is destroyed. There is a kind of ray that will destroy

the tumor if the rays reach it at a high intensity. But unfortunately, such high-intensity rays will also destroy the healthy tissue they pass through on the way to the tumor. At lower intensities the rays will not damage the healthy tissue, but neither will they remove the tumor. How can the doctor use rays to destroy the tumor, while at the same time sparing the healthy tissue?

Solution:

**The doctor could use multiple low-intensity rays from several directions all at once. The rays would converge on the tumor, where their combined effect would be enough to destroy it. Since each spot on the healthy tissue would receive only a low-intensity ray from one source, the healthy tissue would be left intact.**

## S5.4 Solution to radiation problem with distracting stories

The following contains GPT-3's proposed solution to the radiation problem when first presented with both the general story and two distracting (i.e., non-analogous) stories. GPT-3 did not identify the convergence solution in this case, instead proposing the same solution as it did when presented with the radiation problem in isolation.

Story #1 – The Wine Merchants:

One day a rich man found that his wine cellar was empty. So he sent out messengers to announce a generous offer. The first person to bring the rich man a barrel of wine would be given a brick of solid gold. However, the offer would expire at sundown. Two wine merchants heard the news. Each had a horse-drawn cart loaded with large barrels of wine. They both set out for the duke's palace at once. An hour before sundown they came to a place where the bridge had been washed out by a raging river. The first merchant drove his horses and cart into the flood in a desperate attempt to reach the other side. But the horses were already exhausted and could not fight the current. The cart overturned, and the horses, wine, and driver were washed away. The second merchant tried a different tactic. He poured the wine out of all but one of his barrels, and lashed them together to form a raft; then he loaded the one full barrel, a horse, and himself on top. He set the raft adrift and floated downstream. In a few minutes the raft came to rest on the shore in front of the town where the rich man lived. The merchant disembarked, loaded the wine barrel on the horse, and led it to the rich man's house. He arrived just as the sun was setting, and collected the gold brick as a reward for his efforts.

Story #2 – The General:

An evil dictator controlled a fortress situated in the center of a small country. Many roads radiated out from the fortress like spokes on a wheel. A general raised an army at the border, vowing to capture the fortress and overthrow the dictator. The general was about to send his entire army down one road to capture the fortress, when he learned that the dictator had mined each road so that although small groups could still pass, a large army would set off an explosion. The general then had a clever idea: he divided his army into small groups and dispatched each group to the head of a different road. Upon his signal, each group charged down a different road. All the groups passed safely to the fortress, where the entire army attacked the fortress in full strength. In this way the general captured the fortress and overthrew the dictator.

Story #3 – The Identical Twins:

Once there were identical twins who were continually playing pranks on their family, friends, and teachers. The annual school picnic was always a big event for the twins. There were races and other athletic events in which the twins won lots of prizes. One year a new student arrived who was a star runner. The twins wanted to win the main event: the 2-mile race through the woods behind the school. So they secretly devised a plan which would enable them to outdo the newcomer. The day of the race arrived. Each runner was to pick his own path through the woods to a clearing, where a teacher stood posted to

determine the winner. One twin entered the race, while the other excused himself on the grounds that he had hurt his leg in an earlier broadjumping event. The race began and the students rushed into the woods. The twin rushed into the woods and waited until the others had passed out of sight. Then he went back to the school using a path hidden from the picnic area. Shortly after, the other twin, who had been hiding behind a rock near the finish line of the race, burst out and ran into the clearing ahead of the other runners. The teacher named him the winner and marveled at the speed of his running. Next year the twins switched places and thereafter maintained their status on this event.

Target problem:

Suppose you were a doctor faced with a patient suffering from a malignant stomach tumor. The tumor is inoperable, and the patient will die unless it is destroyed. There is a kind of ray that will destroy the tumor if the rays reach it at a high intensity. But unfortunately, such high-intensity rays will also destroy the healthy tissue they pass through on the way to the tumor. At lower intensities the rays will not damage the healthy tissue, but neither will they remove the tumor. How can the doctor use rays to destroy the tumor, while at the same time sparing the healthy tissue?

Solution:

**The doctor can use a technique called brachytherapy. This involves placing a radioactive source directly into the tumor. The radiation from the source will be of a high enough intensity to destroy the tumor, but the radiation will not travel far enough to damage the healthy tissue. The doctor can also use a technique called stereotactic radiotherapy. This involves using a machine to deliver a high-intensity beam of radiation to the tumor from outside the body. The beam is focused on the tumor, so that the healthy tissue is spared.**

## S5.5 Solution to radiation problem with distracting stories and prompt

When presented with distracting stories, GPT-3 did not propose a solution to the radiation problem based on the analogy with the general story. We therefore provided a prompt to encourage GPT-3 to consider the previously presented stories when generating its solution. This same prompt was used in the original human behavioral study,<sup>[22]</sup> where it was found to increase the likelihood that participants would identify the analogous convergence solution. The source stories and target problem were presented in the exact same manner, but the prompt used to elicit GPT-3's solution was changed. When presented with this prompt, GPT-3 correctly identified the convergence solution, and specifically identified the general story as the relevant source.

Solution (in solving this problem you may find that one of the stories you read before will give you a hint for a solution of this problem):

**The doctor could use the same tactic as the general in the second story. He could divide the rays into small groups and send each group down a different path to the tumor. The rays would be of low intensity, so they would not damage the healthy tissue, but when they all reached the tumor at the same time, the combined intensity would be enough to destroy it.**

## S5.6 Solution to gumball problem in isolation

The following contains GPT-3's proposed solution to the gumball problem (from<sup>[23]</sup>) when presented in isolation (enumeration is part of the response provided by GPT-3).

Target problem:

You are seated in front of a table. Two bowls are on the table, one within your reach and one farther away. The closer bowl contains a number of small gumballs, and the farther one is empty. Also on the table are an aluminum walking cane, a large rectangular sheet of heavy paper (posterboard), a hollow

cardboard tube long enough to reach the farther bowl, scissors, string, tape, paper clips, and rubber bands. Using the materials provided, how can you transfer the balls from the filled to the empty bowl without leaving your seat?

Solution:

1. Cut the posterboard into a long strip and tape it to the end of the cardboard tube.
2. Tie one end of the string to the end of the tube and the other end to the handle of the cane.
3. Place the tube in the bowl of gumballs and use the cane to pull the tube and gumballs towards you.
4. Once the tube is close enough, use the paper clips to secure the tube to the edge of the table.
5. Use the rubber bands to secure the tube to the cane.
6. Pull the cane towards you, using the string, and the gumballs will be transferred to the empty bowl.

### S5.7 Solution to gumball problem following magic staff story

The following contains GPT-3's proposed solution to the gumball problem when first presented with the 'magic staff story'. This story involves a magical genie who has moved from his old home (a bottle) to a new home (a different bottle), and wants to transport his collection of jewels between the bottles. To do so, he uses a magic staff to pull the new bottle over next to the old bottle and transfers the jewels by hand. When first presented with this story, children most often identify a solution to the gumball problem that involves using the aluminum walking cane to pull the empty bowl over next to the bowl with the gumballs and then transferring the gumballs by hand. GPT-3 was not able to discover this solution, but was able to identify the high-level analogy between the source story and the target problem (though not the specific analogy between the magic staff and the walking cane).

Source story:

Once upon a time there lived a magical genie. He was a very old, wise, and rich genie indeed. One day while he was polishing his home, which was actually a bottle, he decided he would like to find an even bigger and better home to live in. So he began searching far and wide for another bottle. Finally he found the perfect home. It was larger, prettier, and not too far away from his old bottle. The genie was very excited and began moving his belongings right away. But now the genie had a problem. He had a great many beautiful and very precious jewels in his old home. He had to somehow get all the jewels from his old bottle to the new bottle without dropping or losing a single jewel.

After thinking a bit, the genie came up with a wonderful idea. He began searching for his magic staff, or wand. He then commanded his staff to stretch itself from his old home to his new home. Next, the genie tugged and pulled on his magical staff until at last he pulled the new bottle right up next to his old bottle. At once, the genie began gathering his jewels together in his old home and simply dropped them carefully into his new home right next to him. When all his jewels were safely tucked away in his new home, the genie settled in happily. He invited his friend to come in and admire his new home. I'm sure you can still find the genie sitting in his new bigger and better bottle with all his jewels and smiling contentedly even today!

Target problem:

You are seated in front of a table. Two bowls are on the table, one within your reach and one farther away. The closer bowl contains a number of small gumballs, and the farther one is empty. Also on the

table are an aluminum walking cane, a large rectangular sheet of heavy paper (posterboard), a hollow cardboard tube long enough to reach the farther bowl, scissors, string, tape, paper clips, and rubber bands. Using the materials provided, how can you transfer the balls from the filled to the empty bowl without leaving your seat?

Solution:

**Take the cardboard tube and cut it in half. Place one half of the tube in the bowl with the gumballs and the other half in the empty bowl. Secure the two halves together with the rubber bands. Take the posterboard and cut it into strips. Tape the strips together to form a long rope. Tie one end of the rope to the handle of the walking cane and the other end to the rubber bands. Place the walking cane on the table and hold it in place. Gently pull the cane towards you, and the gumballs will travel through the tube and into the empty bowl. When all the gumballs have been transferred, remove the tube and rope and enjoy your gumballs!**

Can you identify an analogy between the source story and the target problem?

**The source story and the target problem both involve transferring something from one place to another. In the source story, the genie is transferring his jewels from one bottle to another, and in the target problem, the gumballs are being transferred from one bowl to another. In both cases, the characters must use creative problem-solving to figure out how to move the items without losing any.**

Does this analogy suggest a solution to the target problem?

**Yes, the analogy suggests a solution to the target problem. The genie used a magical staff to stretch from one bottle to the other, and the target problem suggests using a rope and a walking cane to transfer the gumballs from one bowl to the other.**

Who or what in the target problem is most analogous to the genie's old home in the source story?

**The bowl with the gumballs is most analogous to the genie's old home in the source story. Both contain something valuable that needs to be transferred to a new location.**

Who or what in the target problem is most analogous to the genie's new home in the source story?

**The empty bowl is most analogous to the genie's new home in the source story. Both are the destination for something valuable that needs to be transferred from an old location.**

Who or what in the target problem is most analogous to the jewels in the source story?

**The gumballs are most analogous to the jewels in the source story. Both are valuable items that need to be transferred from one place to another.**

Who or what in the target problem is most analogous to the magic staff in the source story?

**The rope and walking cane are most analogous to the magic staff in the source story. Both are used to transfer something valuable from one place to another.**

## **S5.8 Solution to gumball problem following magic carpet story**

The following contains GPT-3's proposed solution to the gumball problem when first presented with the 'magic carpet story'. This story has the same setup as the magic staff story – a magical genie wants to transport his jewels

from his old bottle to his new bottle – but the genie’s solution in this story is to roll up his magic carpet and use it as a bridge between the two bottles. When first presented with this story, children most often identify a solution to the gumball problem that involves rolling up the posterboard and rolling the gumballs through it into the empty bowl. It is interesting to note that children do this despite having access to an already formed cardboard tube, likely due to the analogy between the rolled up posterboard and the rolled up magic carpet. GPT-3 identified essentially this same solution, but added some unnecessary steps, including first cutting the posterboard in half. GPT-3 correctly identified the analogy between the source story and the target problem.

Source story:

Once upon a time there lived a magical genie. He was a very old, wise, and rich genie indeed. One day while he was polishing his home, which was actually a bottle, he decided he would like to find an even bigger and better home to live in. So he began searching far and wide for another bottle. Finally he found the perfect home. It was larger, prettier, and not too far away from his old bottle. The genie was very excited and began moving his belongings right away. But now the genie had a problem. He had a great many beautiful and very precious jewels in his old home. He had to somehow get all the jewels from his old bottle to the new bottle without dropping or losing a single jewel.

After thinking a bit, the genie came up with a wonderful idea. He searched for his magic carpet. Then he commanded it to roll itself up into a long hollow tube. Next the genie commanded his flying carpet to place one end at his old home and the other end at his new home so that it formed a sort of hollow bridge between the two bottles. Then, the genie very carefully took one jewel from inside his old home and placed it into the opening of his carpet. At once, the jewel began tumbling and rolling through the carpet tube until it reached his new home and plopped safely inside. The genie grinned happily and began rolling all his jewels through the carpet into his new home. In fact, I’m sure you can still find him sitting in his new, bigger and better bottle with all his jewels and smiling contentedly even today!

Target problem:

You are seated in front of a table. Two bowls are on the table, one within your reach and one farther away. The closer bowl contains a number of small gumballs, and the farther one is empty. Also on the table are an aluminum walking cane, a large rectangular sheet of heavy paper (posterboard), a hollow cardboard tube long enough to reach the farther bowl, scissors, string, tape, paper clips, and rubber bands. Using the materials provided, how can you transfer the balls from the filled to the empty bowl without leaving your seat?

Solution:

**Take the posterboard and cut it into two pieces. Take one of the pieces and roll it up into a tube. Secure the tube with the rubber bands, tape, and paper clips. Place one end of the tube into the bowl with the gumballs and the other end into the empty bowl. Then, use the walking cane to push the gumballs through the tube until they reach the empty bowl.**

Can you identify an analogy between the source story and the target problem?

**The source story and the target problem both involve transferring items from one place to another. In the source story, the genie transfers jewels from one bottle to another using a flying carpet. In the target problem, the person transfers gumballs from one bowl to another using a posterboard tube. In both cases, the items are transferred without leaving the original location.**

Does this analogy suggest a solution to the target problem?

**Yes, the analogy suggests a solution to the target problem. The solution is to use a posterboard tube to transfer the gumballs from one bowl to the other. This is similar to how the genie used a flying carpet to transfer the jewels from one bottle to the other.**

Who or what in the target problem is most analogous to the genie's old home in the source story?

**The bowl with the gumballs is most analogous to the genie's old home in the source story. Both contain items that need to be transferred to a new location.**

Who or what in the target problem is most analogous to the genie's new home in the source story?

**The empty bowl is most analogous to the genie's new home in the source story. Both are the destination for items that need to be transferred from an old location.**

Who or what in the target problem is most analogous to the jewels in the source story?

**The gumballs in the target problem are most analogous to the jewels in the source story. Both are items that need to be transferred from one location to another.**

Who or what in the target problem is most analogous to the magic staff in the source story?

**The posterboard tube in the target problem is most analogous to the magic carpet in the source story. Both are used to transfer items from one location to another.**