

UNIVERSITÀ DEGLI STUDI DI PADOVA

COGNITIVE, BEHAVIOURAL AND SOCIAL DATA

Analogical Reasoning: A comparative analysis between ChatGPT 3.5 and ChatGPT 4

José Chacón, 2071875
Auriane Mahfouz, 2072042

Department of Mathematics

January, 2024

Contents

1	Abstract	2
2	Introduction	2
2.1	Large Language Models (LLMs)	2
2.2	The Debate on LLM Reasoning	2
2.3	Analogical Reasoning with ChatGPT	2
3	Material and methods	3
3.1	The Radiation problem	3
3.1.1	ChatGPT 3.5 and ChatGPT 4 familiarity with the problem	3
3.2	Analogue stories	5
3.3	Methodology	7
3.3.1	Experiment 1	7
3.3.2	Experiment 2	12
4	Results	14
4.1	Experiment 1	14
4.1.1	The Surgeon as base analogue	15
4.1.2	The General as base analogue	17
4.1.3	The Dictator as base analogue	20
4.1.4	The Radiation Problem responses	24
4.2	Experiment 2	27
4.2.1	ChatGPT 3.5	27
4.2.2	ChatGPT 4	31
4.2.3	Results Comparison for Experiment 2	35
5	Discussion	36
6	Conclusion	37

1 Abstract

Large Language Models (LLMs) have emerged as revolutionary tools within the field of cognitive psychology, showcasing astonishing human-like abilities in reasoning, language manipulation, and problem-solving. A prominent representative of these LLMs is GPT, a versatile language model that has drawn considerable attention for its impressive language capabilities. However, one notable challenge that GPT faces is its ability to retrieve analogies, a cognitive task that involves finding connections between seemingly unrelated concepts. This challenge has sparked interest and raised questions about the depth of GPT’s cognitive understanding and its capacity to perform tasks requiring nuanced conceptual associations. The most used application is ChatGPT which is an interface based on GPT. To address this intriguing question and explore the limits of ChatGPT’s cognitive capabilities, this paper draws inspiration from Keane (1987). Keane’s research is renowned for its focus on analogical problem-solving, particularly in the context of remote analogies, which involve identifying connections between domains with minimal attribute overlap. By adapting Keane’s methodology, this paper aims to assess the performance of ChatGPT 3.5 and the current version ChatGPT 4 in retrieving analogies, shedding light on its cognitive abilities in this specific domain.

2 Introduction

2.1 Large Language Models (LLMs)

Large Language Models (LLMs) have emerged as remarkable tools in cognitive psychology, showcasing human-like abilities in reasoning, language manipulation, and problem-solving tasks. With billions to trillions of parameters, these models are trained on extensive text data, predicting the next word in a sequence through self-supervised learning. The evolution of language processing models, particularly from the Perceptron by Rosenblatt (1960) to the Recurrent Neural Network Language Model (RNNLM) developed by Elman (1991), paved the way for the current state-of-the-art Large Language Models (LLMs). The impact of LLMs are owed to two critical advancements: word embeddings and self-attention mechanisms. Self-attention, introduced by Vaswani et al. in 2017 allowed LLMs to capture distant connections within text, reshaping their abilities. This attention mechanism allows models like GPT-4 to comprehend complex problems by identifying crucial associations within text.

2.2 The Debate on LLM Reasoning

In the ongoing debate about Large Language Models (LLMs), critics question their reasoning abilities, while proponents emphasize their impressive language skills. Recent research (Sartori and Orrù (2023)) has shown that LLMs like GPT-3 and GPT-4 can match or surpass human performance in various tasks, such as analogical reasoning and lie detection. However, they struggle with causal reasoning and may produce errors termed “hallucinations.” LLMs improve with practice, fine-tuning, and specific instructions. They align with the theory of associationism in psychology, recognizing associations between concepts and excelling in long-distance reasoning tasks. This challenges previous beliefs about cognitive processes and underscores the importance of understanding and optimizing prompts for LLMs.

2.3 Analogical Reasoning with ChatGPT

Keane (1987) paper presents a model for assessing “analogue retrieval” to study how people retrieve relevant analogies from long-term memory, especially focusing on remote analogies. Remote analogies involve connecting domains with little attribute overlap, while literal analogies have overlapping attributes. The study aims to test the prediction that retrieving remote analogies is challenging due to a lack of semantic cues, exploring the factors influencing this process.

The paper introduces two experiments for studying analogue retrieval and presents initial findings suggesting that remote analogues are indeed challenging to retrieve.

In this study, we aim to replicate and extend the experiments conducted by Keane (1987), with a specific focus on evaluating ChatGPT’s capacity to retrieve analogies. We will adapt the experimental framework and apply it to ChatGPT, allowing us to investigate the model’s performance in this critical cognitive task. Regarding this, we will also compare the performance of ChatGPT 3.5 and ChatGPT 4. By doing so, we seek to shed light on ChatGPT’s abilities, limitations and

improvements in the realm of analogue retrieval, contributing to a better understanding of its cognitive capabilities.

3 Material and methods

In this section, we will present the stories employed in our experiments and provide a detailed description of the methodology used.

3.1 The Radiation problem

This cognitive performance test was created by psychologist Karl Duncker (1945). The problem involves finding a way to destroy a tumor in a patient's body using rays that are harmless at low intensity but harmful at high intensity. This solution can be derived by using analogical thinking, which is the ability to apply knowledge from one domain to another. When the test was introduced, many people fail to notice or use such analogies, due to functional fixedness, which is the tendency to see objects and concepts only in their conventional or familiar roles. The story and its solution are found in Figure 1

Story:

Suppose you are a doctor with a patient who has a malignant tumour in his stomach. It is impossible to operate on the patient; but unless the tumour is destroyed the patient will die. There is a kind of ray that can be used to destroy the tumour. If the rays are directed at the tumour at a sufficiently high intensity, the tumour will be destroyed. Unfortunately, at this intensity the healthy tissue that the rays pass through on the way to the tumour will also be destroyed. At lower intensities the rays are harmless to the healthy tissue but they will not affect the tumour either. What type of procedure might be used to destroy the tumour with the rays, and at the same time avoid destroying the healthy tissue?

Solution:

The ray may be divided into several low-intensity rays, no one of which will destroy the healthy tissue, if these several rays are positioned at different locations around the body and focused on the tumor, their effects will combine and destroy the tumor

Figure 1: The Radiation Problem with its solution

3.1.1 ChatGPT 3.5 and ChatGPT 4 familiarity with the problem

Before collecting data from our conversations with ChatGPT 3.5 and 4, we conducted preliminary conversations with these interfaces to assess the possibility that the internal representation of the models they are based on already had knowledge about Dunckers' Radiation Problem.

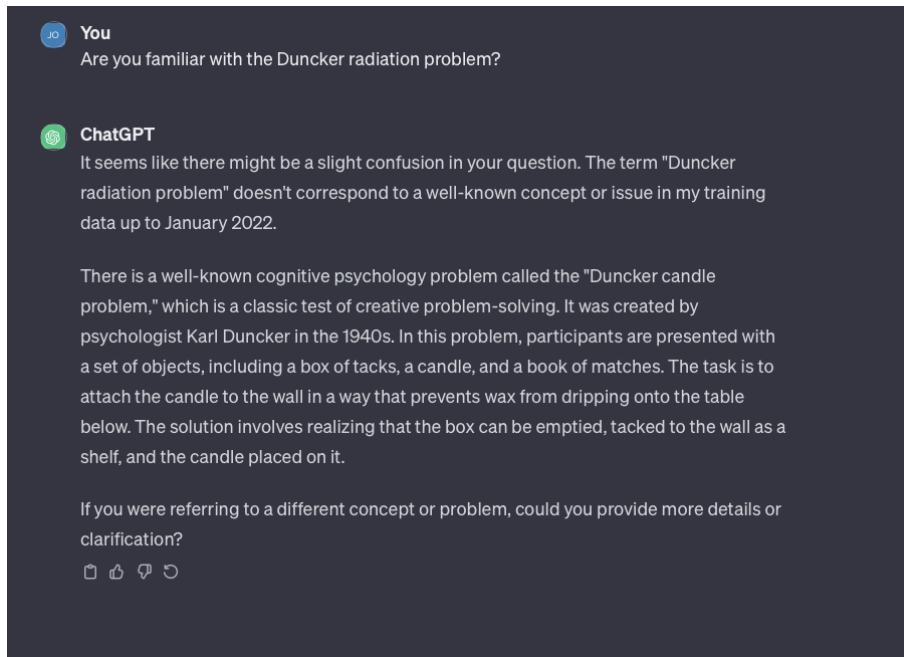


Figure 2: ChatGPT 3 response at the start of the conversation

For ChatGPT 3, when asked "Are you familiar with the Duncker radiation problem?" its response mentioned the current knowledge cut-off, which corresponds to January 2022. According to the response, there was no recognition of the problem. Instead, it mentioned other contribution of the same psychologist, The Candle Problem. Then, the ChatGPT 3.5 proceeded describing it.

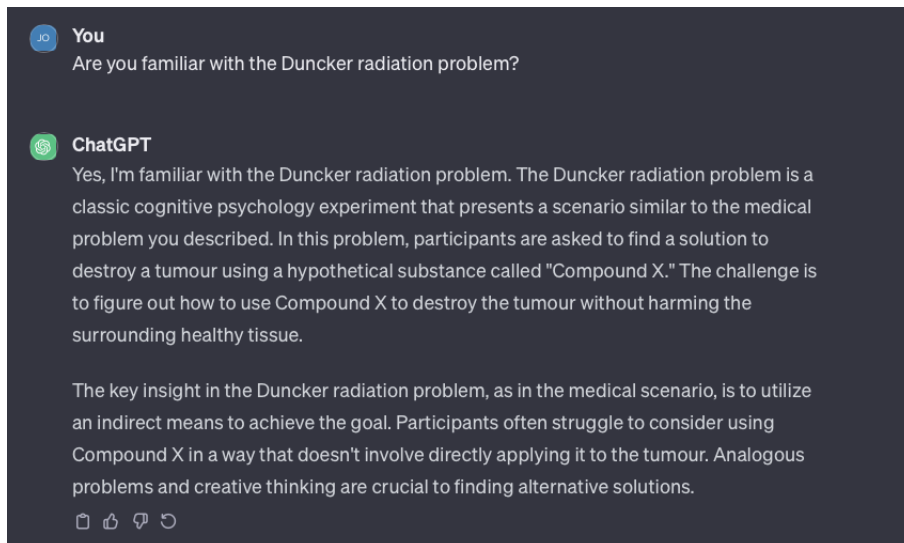


Figure 3: ChatGPT 3 response at the end of the conversation

When we tried a similar prompt in ChatGPT 4, the response acknowledged the model's familiarity with The Radiation Problem. Furthermore, but also detailed the problem, its solution and another variant that is used in cognitive experiments to assess Analogical Reasoning. This particular variant, "The General", is used in this experiment as it appeared in the experiments from which this work draws influence (Keane (1987))

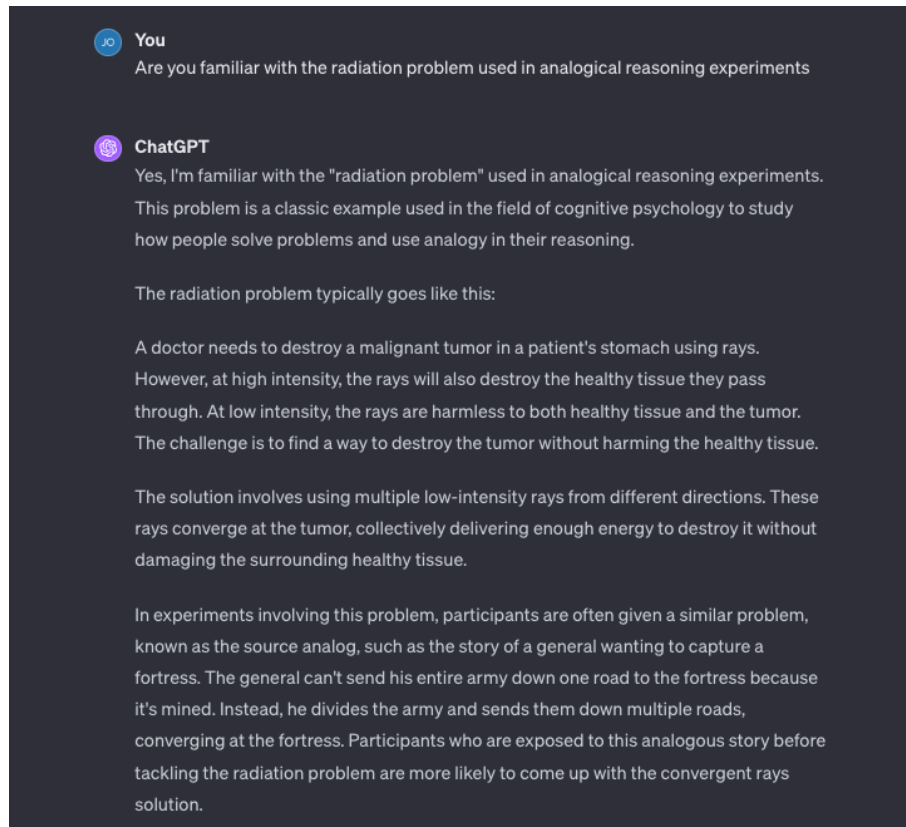


Figure 4: ChatGPT 4 response

3.2 Analogue stories

A brief description of the five analogue stories is given.

1. **The Surgeon:** This story was introduced by Gick and Holyoak (1980). The elements of the story resembled considerably The Radiation Problem. The goal is similar, but with slightly changes in other features of the story such as the tumor location which now is situated in the patient's brain. The story is found in Figure 5.

A surgeon was trying to destroy a cancer which was situated in the central region of a patient's brain, by using a type of ray. He needed to use these rays at a high intensity in order to destroy the cancerous tissue. However, at such an intensity the healthy brain tissue will also be destroyed. After considerable thought, he knew just what to do. He divided the rays up into batches of low-intensity rays, and by sending them, simultaneously, from a number of different directions, they converged on the cancer, making up a sufficiently high intensity to destroy it.

Figure 5: The Surgeon story

2. **The General:** This story was introduced by Gick and Holyoak (1980) as The Fortress Story. Although in this work, we're using the adaptation presented in Keane (1987) The analogy between The General story and The Radiation Problem is located in the shared features between the fortress and the tumor. The small groups of soldiers sent down different roads in the story correspond to the low-intensity rays that can be directed at the tumor in the radiation problem. The story is found in Figure 6.
3. **The Dictator:** This story is an adaptation of "The Fortress Story" appearing in Gick and Holyoak (1980). There has been slight modifications in order to be aligned with the model of analogue retrieval introduced in Keane (1987). This was the addition of a goal action ("destroy") semantically close to The Radiation Problem goal. The story is found in Figure 7.

A general was trying to destroy a fortress which was situated at the centre of a country with roads leading to it, by using his army. He needed to use his army as a complete group in order to destroy the fortress. However, he could not march his army down a road to the fortress because the roads were mined to explode when large groups of men passed over them. After considerable thought, he knew just what to do. He divided his army up into small groups of men, and by sending these groups, simultaneously, from a number of different directions, they converged on the fortress, making up a sufficiently powerful army to destroy it.

Figure 6: The General story

A small country fell under the iron rule of a dictator. The dictator ruled the country from a strong fortress. The fortress was situated in the middle of the country, surrounded by farms and villages. Many roads radiated outward from the fortress like spokes on a wheel. A great general arose who raised a large army at the border and vowed to destroy the fortress and free the country of the dictator. The general knew that if his entire army could attack the fortress at once it could be destroyed. His troops were posed at the head of one of the roads leading to the fortress, ready to attack. However, a spy brought the general a disturbing report. The ruthless dictator had planted mines on each of the roads. The mines were set so that small bodies of men could pass over them safely, since the dictator needed to be able to move troops and workers to and from the fortress. However, any large force would detonate the mines. Not only would this blow up the road and render it impassable, but the dictator would then destroy many villages in retaliation. A full-scale direct attack on the fortress therefore appeared impossible. The general, however, knew what to do. He divided his army up into small groups and dispatched each group to the head of a different road. When all was ready he gave the signal, and each group marched down a different road. Each group continued down its road to the fortress, so that entire army finally arrived together at the fortress at the same time. In this way, the general was able to destroy the fortress, and thus capture the dictator.

Figure 7: The Dictator story

4. **General-Ray:** Variant of The General story as it appeared in Keane (1987). Here, the army was replaced by the use of rays. With this change, the instrument used is identical to The Radiation Problem. The story is found in Figure 8.
5. **General-Laser:** Variant of The General story as it appeared in Keane (1987). The instrument mentioned in the story is laser which is semantically close to the rays used in The Radiation Problem. The story is found in Figure 9.

A general was trying to destroy an intercontinental ballistic missile, using a type of ray, which would be entering the atmosphere at great speeds. He needed to use a high-intensity ray in order to destroy the incoming missile. However, such high-intensity rays were inaccurate because they would heat up the air through which they passed, thus distorting the beam. After considerable thought, he knew what to do. He divided the high-intensity rays into a number of low-intensity rays and, by sending these low-intensity rays, simultaneously, from a number of different directions, they converged on the missile making up a sufficiently powerful beam to destroy it.

Figure 8: The General-Ray story

A general was trying to destroy an intercontinental ballistic missile, using a type of laser beam, which would be entering the atmosphere at great speeds. He needed to use a high-powered laser beam in order to destroy the incoming missile. However, such high-powered laser beams were inaccurate because they would heat up the air through which they passed, thus distorting the beam. After considerable thought, he knew what to do. He divided the high-powered laser beams into a number of low-powered laser beams and, by sending these low-powered laser beams, simultaneously, from a number of different directions, they converged on the missile making up a sufficiently powerful beam to destroy it.

Figure 9: The General-Laser story

3.3 Methodology

3.3.1 Experiment 1

Following the methodology detailed in the initial experiment by Keane (1987), the conversations with ChatGPT 3.5 and 4 were divided in two stages: Acquisition Stage and Test Stage.

During the Acquisition Stage, the LLM is asked to analyse the story and recall it precisely. In the Test Stage, the model’s analogue retrieval ability is assessed by requiring it to solve the Radiation Problem. The expected outcome is that the LLM would retrieve the correct analogue that corresponds to the story analogue used, by acknowledging the similarities in its response. Otherwise, the focus of the conversation would shift towards enforcing the mapping of the similarities between the base analogues and the target problem.

The stories used in the experiment were The Surgeon, The General and The Dictator as base analogue. The target problem was the Duncker’s Radiation Problem. The prompts used in the experiment has been carefully designed following the instructions in Keane (1987), provided to subjects participating in its experiments. For example, the subjects were instructed to think of analogous problems and also reminded to verbalize anything that came to their mind while reading the target problem before trying to solve "The Radiation Problem". These instructions are convenient as they could easily be adapted into a Zero-Shot Chain of Thought prompting type , which has been reported to guarantee an improved performance during a cognitive testing as shown in Kojima, Gu, Reid, Matsuo, and Iwasawa (2022).

Motivated by the recent findings on the cognitive abilities of LLMs, we wanted to study two cases for our experiment:

- Simple case: The successful scenario is represented through two exchanges between user and ChatGPT. One for the Acquisition stage and the other evaluating the LLM ability to solve the Radiation Problem by retrieving the analogue (Test Stage).
- Full case: Following the Acquisition Stage, the LLM is tasked with two problem solving tasks before introducing "The Radiation Problem". Overall, this conversation structure mirrors the process applied on the Test Stage in Keane (1987), where subjects are encouraged to work on problem solving tasks based on analogies.

The reason why both scenarios were chosen is because we wanted to evaluate what difference it would make to have a portion of the conversation separating the introduction of the base analogue and the target problem. While both ChatGPT 3.5 and ChatGPT 4 are known for being capable to decode long-distance relationships, this approach would give us more evidence of the improvements found in ChatGPT 4 compared to its previous version. The workflow representative of each case is detailed in Table 1 .

	Simple case
1	Introduce the base analogue and ask for analysis and recall
2	Induce the LLM to think of analogues. After that,ask the radiation problem
	Full case
1	Introduce the base analogue and ask for analysis and recall
2	Ask analogical reasoning problem number 1
3	Ask analogical reasoning problem number 2
4	Induce the LLM to think of analogues. After that,ask the radiation problem

Table 1: Proposed main workflow for Experiment 1

3.3.1.1 Prompt design

As noted in the preceding section, the prompts design is largely inspired by the instructions given to subjects participating in experiments by Keane (1987)

While the acquisition stage in Keane (1987) took around 5 minutes for participants to memorize the story analogue and 30 minutes to recall it, considering the specific context of our experiment, the prompt for this stage in our experiment is intended to begin in a friendly manner and then transition to this stage goal. The start of the prompt before introducing the story analogue is the following:

Hello, I'll give you a story. I need you to analyse it carefully and then recall it, as precisely as possible. The story is the following: [...]

Proceeding with the Test Stage, in the simple case, the subsequent exchange between user and LLM is of paramount importance in the conversation. The central concept revolves around the prompt's role in steering the LLM's internal representation in the right direction, facilitating an associative connection to decode the analogue story and the Radiation Problem. We found the instructions used in Keane (1987) to be concise and precise for our needs, which is to give enough information to guide the model towards the analogue retrieval. Those are highlighted in the following frame as they appeared in the full prompt:

Before attempting to solve the following task try to think of analogous problems. Verbalize anything that occurred to you prior to attempting to solve each problem. Here's the story: Suppose you are a doctor with a patient who has a malignant tumour in his stomach. It is impossible to operate on the patient; but unless the tumour is destroyed the patient will die. There is a kind of ray that can be used to destroy the tumour. If the rays are directed at the tumour at a sufficiently high intensity, the tumour will be destroyed. Unfortunately, at this intensity the healthy tissue that the rays pass through on the way to the tumour will also be destroyed. At lower intensities the rays are harmless to the healthy tissue but they will not affect the tumour either. What type of procedure might be used to destroy the tumour with the rays, and at the same time avoid destroying the healthy tissue?

For the full case scenario, it was decided to use two problem solving tasks, which were identified in Sartori and Orrù (2023), preceding the part of the conversation where the Radiation Problem is given to ChatGPT to solve. The chosen tasks go as follows:

Problem 1: Here is a bag filled with popcorn. There is no chocolate in the bag. The bag is made of transparent plastic, so you can see what is inside. Yet, the label on the bag says 'chocolate' and not 'popcorn.' Sam finds the bag. She had never seen the bag before. Sam reads the label. She believes that the bag is full of (...)

Problem 2: Mr. and Mrs. Plum have six daughters and each daughter has one brother. How many people are in the Plum family?

In the same sequence as presented, each of these problems was asked to ChatGPT. These prompts were structured as follows:

Prompt 1: Now I'll give you three problem solving tasks. This is the first problem: [...]

Prompt 2: Great, here comes the second problem: [...]

After that, the flow of the conversation is the same as it was stated before for the simple case scenario. The Radiation problem is introduced and the main workflow of the experiment is completed. In both cases, if ChatGPT is not capable of retrieving the right analogue then the flow

of the conversation shifts to receive more information about what caused the model to not behave as it was expected and also to enforce it to map the similarities by explicitly mentioning the base analogue. For this alternative workflow, the inclusion of the following prompts in our conversations would depend on ChatGPT responses :

Prompts for inspecting model behaviour

- How did you arrive at this solution?
- Reflect carefully again

Prompt for enforcing mapping Have you seen a similar story?

Prompt for inspecting model behaviour after mapping: Did you think of using the story to solve the problem?

Prompts to confirm successful mapping:

- Try to solve the problem using it
- Could you now use it to solve the problem
- Recall the analogue

3.3.1.2 Retrieval and Mapping measure

Now that the prompts designed to evaluate ChatGPT 3.5 and 4 abilities has been presented, a matter of interest is to define what constitutes as retrieval and mapping.

Following the methodology outlined in Keane (1987), there are two criteria for the ChatGPT response to be qualified as a successful retrieval. The primary measure involves observing features common to both the base analogue and the target problem in the model's response when asked to solve the Radiation Problem. For instance, this may include shared elements such as the common use of rays and the shared goal of destroying a tumor while avoiding harm to healthy tissue.

The secondary measure of retrieval could be detected after a question such as *How did you arrive at this solution?* is asked. This serves as a second opportunity for ChatGPT to align its response with shared goals and features, providing a better explanation of its initial response.

The model of analogue retrieval in Keane (1987) conceived taking into account the two types of features presented in both base analogue and target problem, the ones related to the conflicted goals and to the objects appearing in the stories. These are a valuable tool to evaluate the responses generated by ChatGPT, therefore the Table 2 is an adaptation of the framework presented in Keane (1987), with the contribution of incorporating The Dictator story along with its respective features.

Story	Goal 1 feature	Goal 2 feature	Object features	
The Radiation Problem	Destroy Object 1	Avoid destroying Object 2	Doctor (actor) Ray/High (instrument 1a) Ray/Low (instrument 1b) Tumour (object 1) Healthy Tissue (object 2)	
Story	Goal 1 feature	Goal 2 feature	Object features	Indices
The Surgeon	Destroy: Object 1*	Avoid destroying: Object 2*	Surgeon (actor)* Ray/High (instrument 1a)* Ray/Low (instrument 1b)* Cancer (object 1)* Healthy Tissue (object 2)*	7
The General	Destroy: Object 1*	Avoid destroying: Instrument 1a	General (actor) Army/Whole (instrument 1a) Army/Part (instrument 1b) Fortress (object 1)	1
The Dictator	Destroy: Object 1* Capture: Actor 2	Avoid destroying: Instrument 1a Avoid destroying: Object 2	General (actor 1) Dictator (actor 2) Army/Whole (instrument 1a) Army/Part (instrument 1b) Fortress (object 1) Village (object 2)	1

Table 2: Goal and Object Features of the Materials in Experiment 1 and 2

3.3.2 Experiment 2

In Keane (1987), Experiment 1 provided limited insight into the processes underlying analogue retrieval. Therefore, he wanted to test the hypothesis that literal analogues, which contain identical elements, are retrieved more often.

Following Keane’s methodology for experiment 2, we will focus on exploring how ChatGPT performs in retrieving analogies using identical words between the analogue story and the problem while comparing both version 3.5 and version 4. For that, starting from "The General" story, two new story analogues were created: The General-Ray and the General-Laser. These analogues were nearly identical, differing only in the instrument used in one ("rays") being identical to the problem, while in the other ("lasers"), it was semantically close but not identical.

The second experiment was also divided into two stages: Acquisition Stage and Test Stage, which were identical to that of Experiment 1. Our approach involves a simple case and a full case. This is useful to evaluate how effectively ChatGPT can utilize information from previous interactions to solve subsequent problems. This division aims to assess ChatGPT’s capacity for retaining and applying past conversational data. By doing so, we seek to understand the extent to which ChatGPT can build upon earlier dialogues to provide more contextual and informed responses.

- Simple case: The LLM is asked to recall the story then solve the problem directly. If it did not retrieve the analogy, it is asked again to think of similar problems.
- Full case: Three problems are asked before trying to solve the radiation problem. Compared to the previous experiment, we wanted to explore how the LLM performs with three problems instead of two.

For each case, a three-scenarios process is presented. However, if ChatGPT successfully passes the first stage, it indicates that it excels at retrieving the analogy directly. We will start with the acquisition stage when at first ChatGPT is given the story and asked to recall it, then in the test stage, asked to solve the radiation problem with altered prompts. The aim of employing diverse prompts is to evaluate ChatGPT’s ability to actively recognize contextual analogies and retain information from a conversation, even when we introduce distractions by posing unrelated questions.

- First scenario: ChatGPT is asked to solve the problem without telling it to think of an analogous story.
- Second scenario: ChatGPT has failed to retrieve the analogy in the first stage, and is therefore asked to solve the problem while mentioning to think of an analogous problem.
- Third scenario: ChatGPT did not retrieve the analogy in the previous two steps and is asked to reflect carefully again.

3.3.2.1 Prompt design

Experiment 2 will be also divided into two stages: Acquisition Stage and Test Stage. The acquisition stage was similar to that of Experiment 1 (section 3.3.1.1), therefore, the same prompts were used. Regarding the test stage, we will conduct separate tests for ChatGPT 3.5 and ChatGPT 4. We will assess each model’s performance in retrieving analogies, first implicitly and then through an explicit question.

Before asking to solve the radiation problem, implicitly or explicitly, a series of 3 problems are asked to the model and the same prompt design was followed to enforce the mapping.

Problem 1: Here is a bag filled with popcorn. There is no chocolate in the bag. The bag is made of transparent plastic, so you can see what is inside. Yet, the label on the bag says ‘chocolate’ and not ‘popcorn.’ Sam finds the bag. She had never seen the bag before. Sam reads the label. She believes that the bag is full of (...)

Problem 2: John pulled out a bag of purple metallic keys, a square hardcover book, a green round melon, a red ripe banana, a sparkling purple gem, a square red card, and a fuzzy purple hat. He placed the first red item he pulled out into his bedroom safe, the second purple item into his fruit bowl and everything else on his living room table. What did John place into his fruit bowl?

Problem 3: What is the first name of the father of Sebastian’s children?

Figure 10: The three problem solving tasks

3.3.2.2 Retrieval and Mapping measure

Experiment 2 builds upon the evaluative framework of Experiment 1. This means that we are going to use retrieval and mapping to measure the performance of ChatGPT. Analogue stories are mapped when retrieval was not possible, thus we induce the model to map the similarity between both stories.

4 Results

4.1 Experiment 1

For both cases, ChatGPT responses were analyzed, and each conversation was tagged as either a successful retrieval or a successful mapping. In contrast to the experiments detailed in Keane (1987), there were no conversations where the response to "The Radiation Problem" was incorrect. This should not come as a surprise, given that the language models (LLMs) these interfaces are based on are trained on a vast internet dataset. This fact becomes evident when the responses mention specialized techniques used in medicine (e.g Gamma Knife surgery, Stereotactic Body Radiation Therapy).

More information on the obtained results and behaviors found will be detailed in the subsections corresponding to each base analogue story used. Overall, in Table 3, the results that we obtained while measuring retrieval or mapping are shown.

Panel A: Simple case			
Story	Condition	GPT 3.5	GPT 4
The Surgeon	Literal Analogue	Retrieved	Retrieved
The General	Remote Analogue	Mapped	Retrieved
The Dictator	Remote Analogue	Mapped	Retrieved

Panel B: Full case			
Story	Condition	GPT 3.5	GPT 4
The Surgeon	Literal Analogue	Retrieved	Retrieved
The General	Remote Analogue	Mapped	Retrieved
The Dictator	Remote Analogue	Mapped	Retrieved*

* Retrieval of the base analogue obtained after a change in the prompt

Table 3: Results for Experiment 1

4.1.1 The Surgeon as base analogue

As presented in Table 3 , since The Surgeon story serves as a literal analogue of The Radiation Problem, the results aligned with our expectations.

ChatGPT-4 successfully retrieved the analogues between both stories. In both the simple and full case scenarios, there is minimal difference in the quality of the responses. From the start of the conversation , the similarities are pointed out. One of the observed responses is found in Figure 11

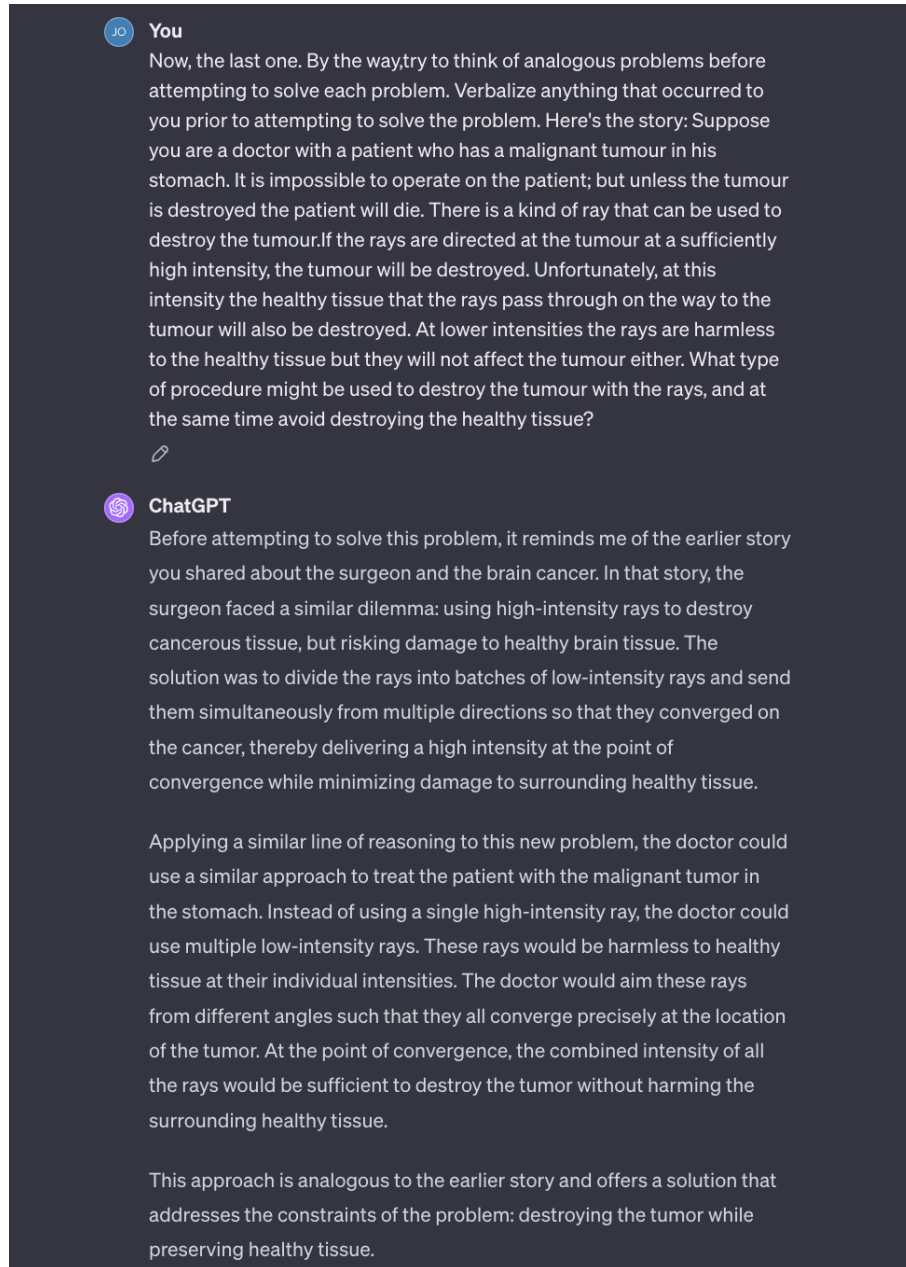


Figure 11: ChatGPT-4 Successful retrieval when the base analogue is The Surgeon story

Similarly, when using ChatGPT-3.5, the same pattern was found. Although we found that in some cases the response that decode the retrieval of the story analogue is found in the very end of the response. However, it is noteworthy that in some instances, the response that decodes the retrieval of the story analogue is located towards the end of the reply. This suggests a potential indication that the internal representation of the model may not have initially recognized the analogue. Supporting this assertion is the fact that other analogies are proposed before suggesting a solution to the target problem. The section of the response where the retrieval occurs is expressed as follows: *'Considering analogous problems and the solution presented in the initial story about brain cancer, a similar approach can be applied.'* The explained behaviour is shown in Figure 12

For sure, both models used different phrases. The ones used in the ChatGPT-4 to retrieve the analogues were: *The problem described in the story is closely analogous to the previous story about*

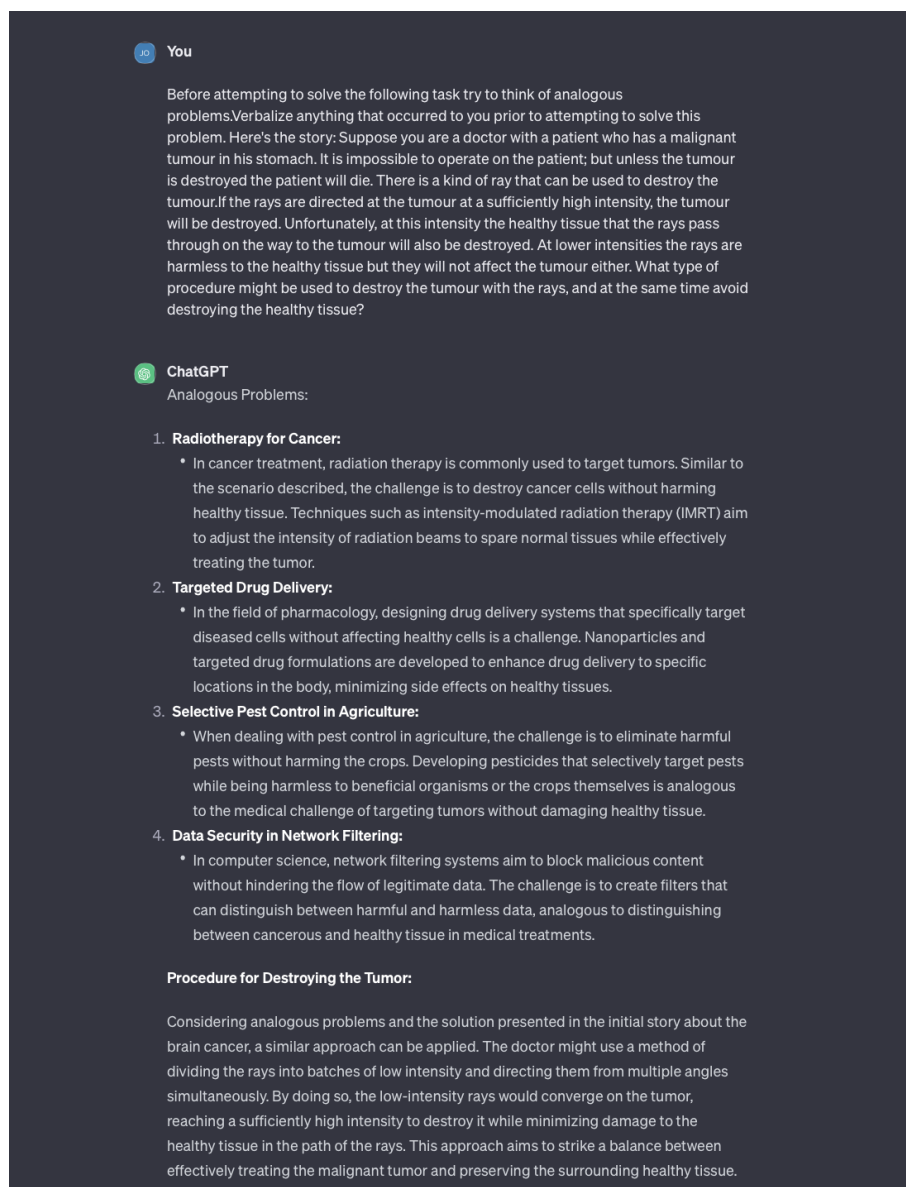


Figure 12: ChatGPT-3 Successful retrieval when the base analogue is The Surgeon story

treating a brain tumor with rays, when faced with the problem of destroying a malignant tumor in the stomach using rays. I would consider the analogous problem presented in the previous story about a surgeon dealing with a brain cancer.

4.1.2 The General as base analogue

As "The General" is a remote analogue to The Radiation Problem, the results aligned closely with what Keane observed in his experiments (Keane (1987)). Specifically, only 12 % of subjects were able to successfully retrieve the analogues.

Through our preliminary interactions with ChatGPT 4, as outlined in section 3.1.1, we had already discovered that ChatGPT 4 is familiar with "The General" story. Therefore, it was not surprising that it could promptly retrieve the analogue from the beginning of its response (Figure 13)

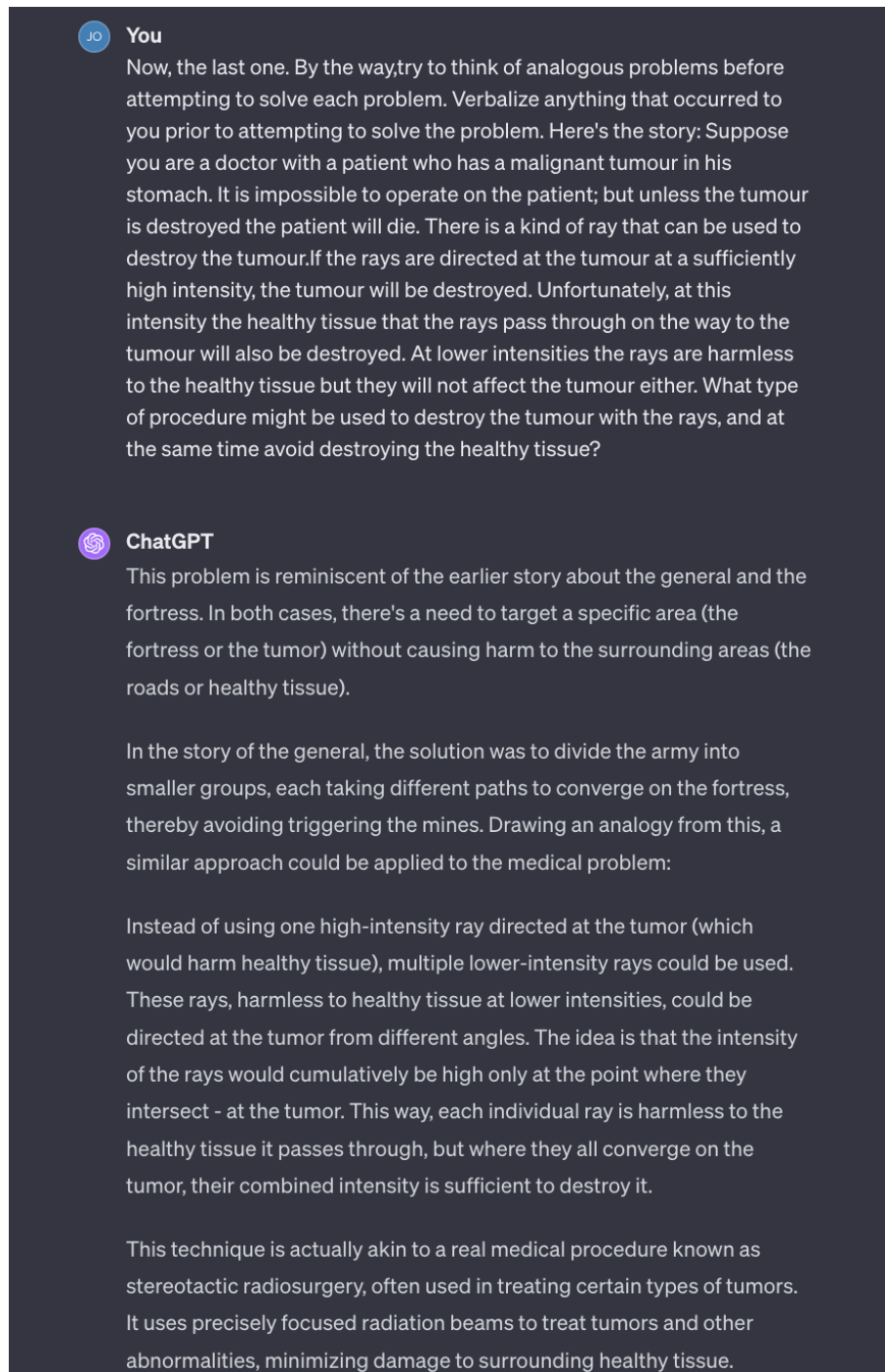


Figure 13: GPT-4 Retrieval when the base analogue is The General story

We could not say the same for ChatGPT 3.5. There was no conversation where it could retrieve the analogue, not even in the simple case where the relationships between stories could be deciphered easily due to their closeness in the flow of the conversation. Therefore, our approach was limited to enforcing the model to map the similarities between both stories. However, there was one instance where the proposed analogue in its response was very similar to what we were expecting. This response could be seen in Figure 14

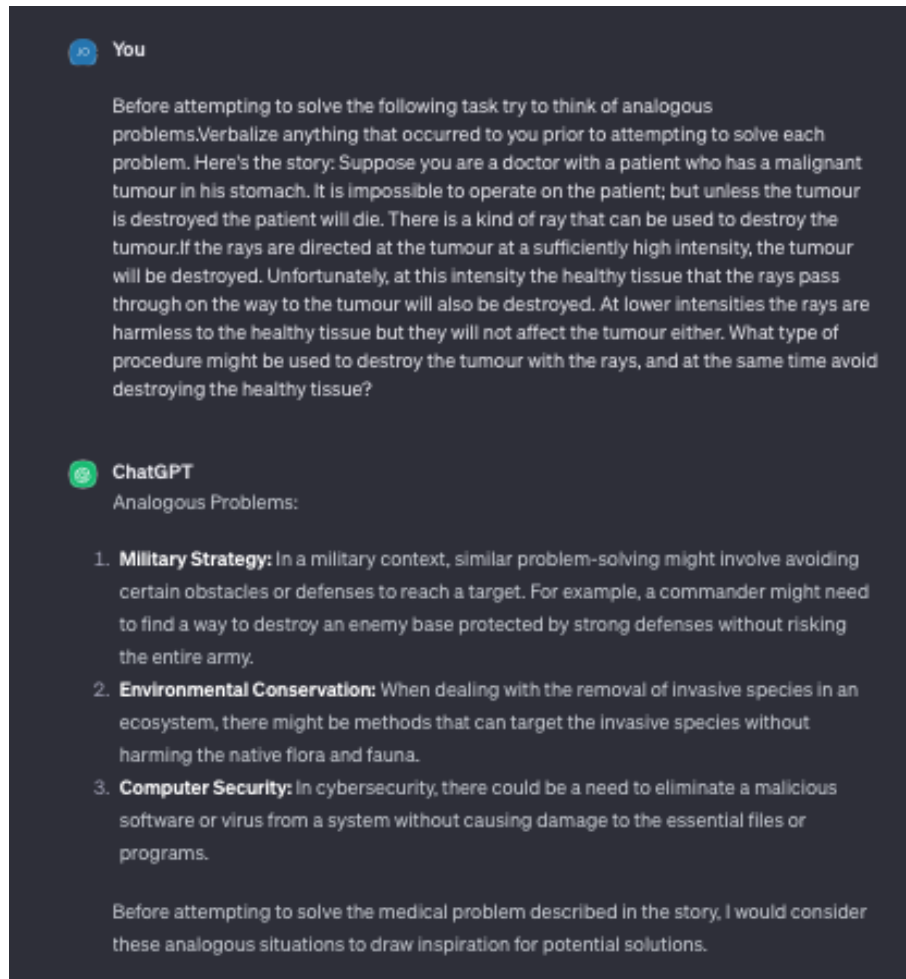


Figure 14: GPT-3.5 response when the base analogue is The General story

In this conversation, the first suggested analogous problem was a scenario where it's necessary to avoid obstacles or defenses to reach a target. This example shares similarities with the goals found in "The General" story. Subsequently, we asked ChatGPT 3.5 *"How did you arrive at this solution?"*, mirroring the approach used in Keane (1987) as a secondary retrieval measure. Keane employed this measure when subjects solved the Radiation Problem but did not explicitly express the analogue with the story presented in the seminar that occurred three days before the experiment. The response to that prompt is shown in Figure 15

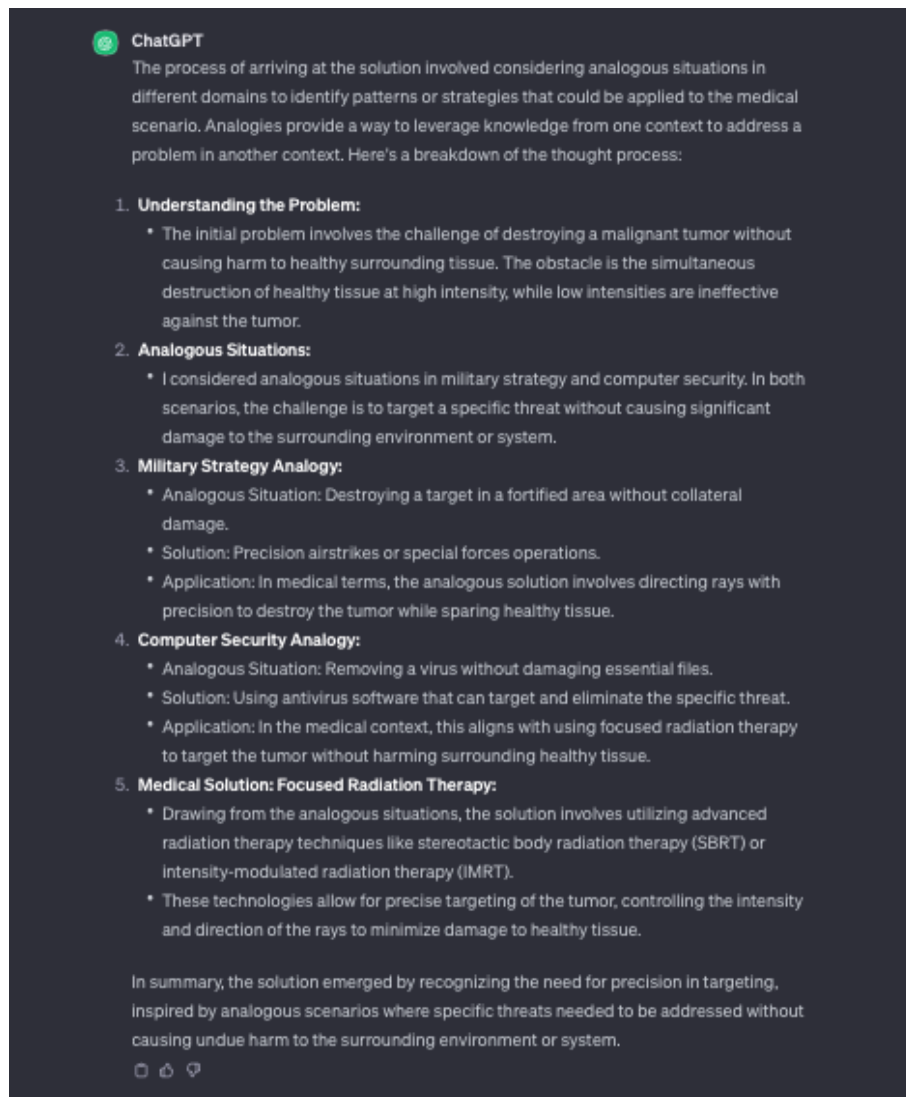


Figure 15: GPT-3.5 reasoning in its Radiation Problem solution when the base analogue is The General story

Let's take a better look to what it said related to its reasoning behind the military strategy analogy:

Military Strategy Analogy:

Analogous Situation: Destroying a target in a fortified area without collateral damage.

Solution: Precision airstrikes or special forces operations.

Application: In medical terms, the analogous solution involves directing rays with precision to destroy the tumor while sparing healthy tissue.

While it gave us a step by step process in its reasoning behind using the military strategy to approach the Radiation Problem, there was not any mention to the base analogue. Therefore, this particular test was not marked as a secondary retrieval.

4.1.3 The Dictator as base analogue

"The Dictator" story follows a similar theme as "The General" story. The only feature that resembles the target problem is the goal of destroying an object, in this case, the fortress, while having the secondary goal which is capturing the dictator. As explained in the previous section, our initial model, ChatGPT 3.5, struggled to decipher the similarities between stories. This challenge persisted when testing it with 'The Dictator' as the base analogue for both the simple and full case scenarios.

As we confirmed that the retrieval would not be possible to be obtained, there was no other way to rely on the prompts we designed for the alternative scenario. So we explicitly asked the model to solve the solution by using "The Dictator" story. By enforcing the mapping of analogues, ChatGPT 3.5 successfully decoded the analogue, expressing it as follows: "In the medical context, the 'mines' are represented by the healthy tissue that can be harmed by the rays."

In another conversation, when asked, "Did you think of using it in the problem?" the model responded: "While the general's strategic approach in dividing the army into small groups to bypass the mined roads shares a concept of strategic division and coordination, the specific details of the story did not directly apply to the medical scenario." The response emphasized the semantic closeness between the features of the target problem and other procedures, such as precision targeting. We interpreted it as a way of defending its decision to recommend precision targeting as an analogue for the target problem.

Most of our findings for "The Dictator" story occurred during the testing of our prompts on ChatGPT 4.0. First, we will comment on its response for the simple case. The way ChatGPT 4.0's response retrieved the analogues before solving the problem is presented in the following frame::

This problem brings to mind an analogous situation from the earlier story about the general and the dictator's fortress. In that story, a large force could not directly attack the fortress due to mines, but small groups could approach safely from different roads. The solution was to divide the large force into smaller groups, each traveling a different path to the fortress, converging simultaneously.

After receiving a satisfactory response, we continued testing the full case in a new conversation. The prompts were the same as before, but now introducing two problem solving questions between the introduction of the base analogue and the introduction of "The Radiation Problem", representing the full case scenario.

To our surprise, the response this time was very similar to the one obtained by using "The General" story as base analogue. For example the proposed analogues were related to targeted drug delivery in cancer treatment, a system for pest control in agriculture and targeted therapy. Similar to other situations explored before, one interpretation could be that the internal representation of the model decided to rely on the vast amount of knowledge it has been trained on instead of decoding the relationships between both stories. The full response is shown in Figure 16

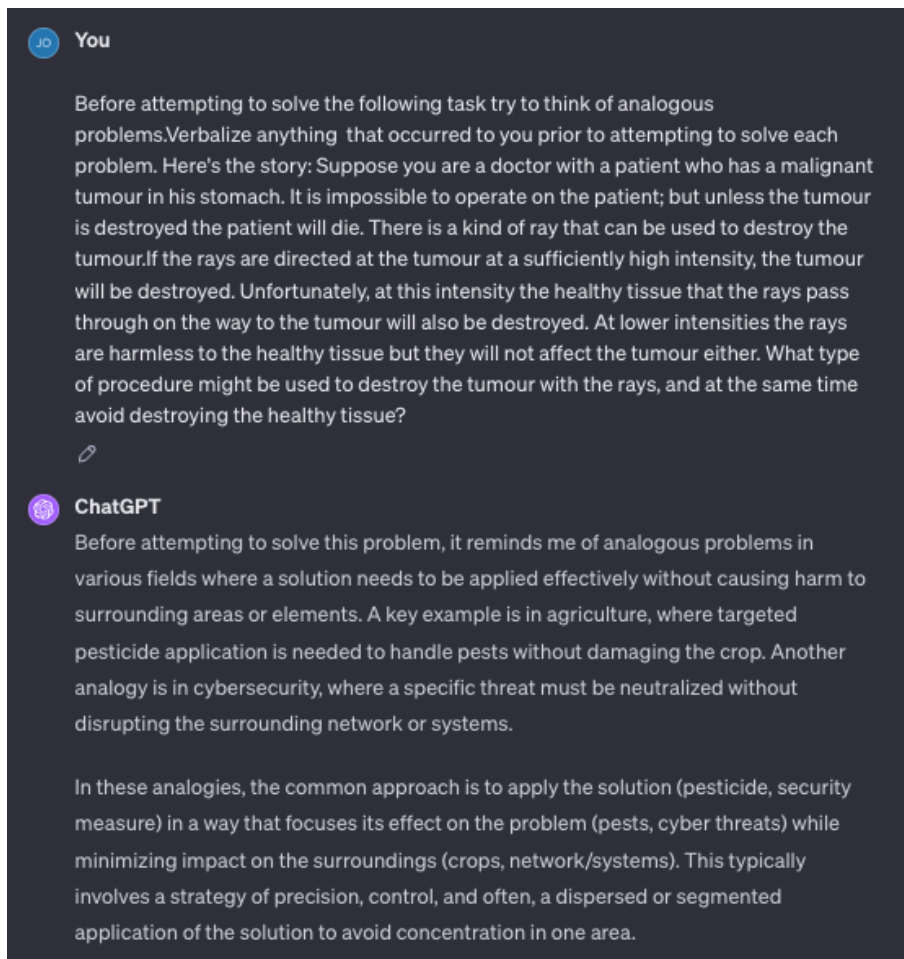


Figure 16: GPT-4 being unable to retrieve the analogue (Full case scenario)

That was not the only unexpected response. As it was explained in Section 3.3.1.1, one of our prompts used to enforce the mapping is *"Have you seen a similar story?"*. While in the previous base stories, this prompt didn't lead to a mapping by the model. This time, we got our first mention of "The Radiation Problem" in our conversations with ChatGPT. We already were aware that ChatGPT 4.0 was somewhat familiar with both "The Radiation Problem" and "The General" but we weren't expecting the model to directly reference the target problem itself. The response we are referring to is shown in the following frame:

Yes, the scenario you described is a classic problem often used in discussions about lateral thinking and creative problem-solving. It's known as the "radiation problem" and was popularized by Edward de Bono, a psychologist and author known for his work in lateral thinking.

In the original radiation problem, the challenge is to destroy a malignant tumor using rays that are too powerful for healthy tissue to withstand. The solution, as discussed, involves using multiple low-intensity rays from different angles, converging at the tumor, thereby destroying it without damaging the surrounding healthy tissue. This problem is typically used to illustrate the concept of thinking creatively and finding innovative solutions to complex problems where direct approaches might be harmful or ineffective.

The similarity between the scenario you presented and the classic radiation problem lies in the need for a solution that is both effective (in destroying the tumor) and precise (to avoid harming healthy tissue), which is a common theme in many lateral thinking puzzles and real-world problems.

After thoroughly reviewing this response, we encountered an unexpected name: *"It's known as the 'radiation problem' and was popularized by Edward de Bono, a psychologist and author known for his work in lateral thinking"*. Both of us are not familiar with Edward de Bono's work.

After an exhaustive search we could not find evidence that de Bono popularized or even used the Duncker's Radiation Problem in his famous book "Lateral Thinking". This could be seen as an example of hallucination in LLMs which has been previously documented on (reference) . Because the primary focus in this work was not studying hallucinations in LLMs , we continued with our tests on "The Dictator story".

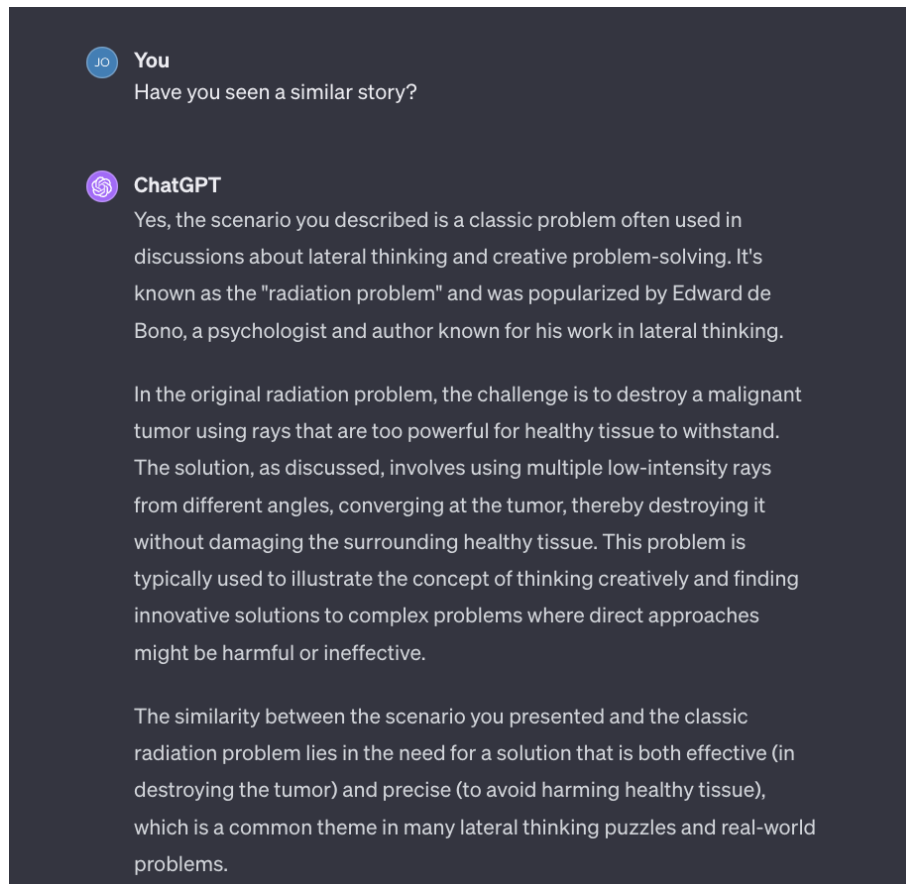


Figure 17: ChatGPT4 - Possible sign of hallucination

Afterward, we directed our focus towards finding what was the cause of the model not retrieving the analogues as expected. We started removing the problem solving tasks that precedes the introduction of "The Radiation problem" but that didn't have any effect on the response. We did that because it's believed that LLMs has certain windows of context as explored in (reference) ,even though we were using one of the most powerful LLM at the moment. We also tried to include all our prompts in one huge message but that also didn't help. In the end, we tried modifying the part of the prompt where we induced the model to make known its reasoning behind its solution to the problem. By trial and error, we ended up finding a prompt that gave us the analogue retrieval. In the following frame, both the original part of the prompt and its modified version are presented:

Original part of the prompt : Before attempting to solve the following task try to think of analogous problems.Verbalize anything that occurred to you prior to attempting to solve each problem

Modified part of the prompt: Verbalize any analogues that occurred to you prior to attempting to solve this problem

With that modification in the prompt, we got a successful retrieval. The full response appears in Figure 18

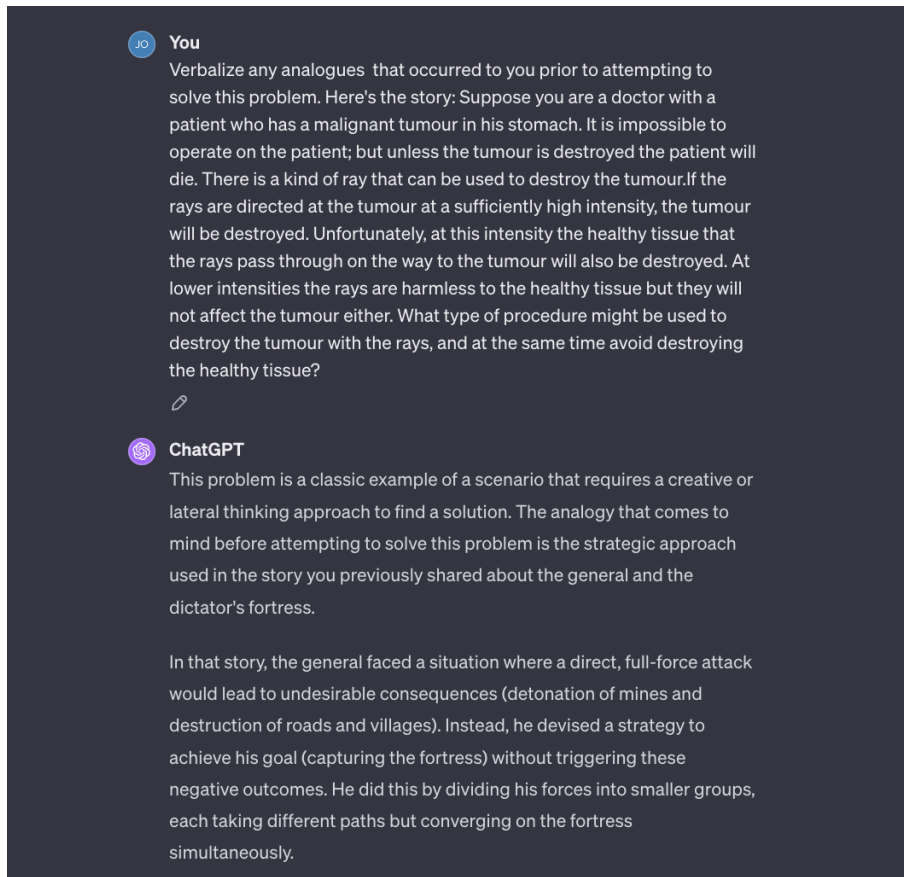


Figure 18: A simplification of the prompt lead to a succesful retrieval

We took advantage of ChatGPT 4.0 not retrieving originally the analogues, to explore other prompts to make the model reevaluate their answers. As simple as it seems, we inputted the considered magic phrase "Reflect carefully again" three times when the model failed to retrieve the base analogue. The responses that we got and specially the new proposed analogues suggested were even more remote than The Dictator Problem itself. While we were familiar with some of these analogues from our previous tests, we discovered new ones, such as the forced analogue between the target problem and a targeted financial investment scenario. The reasoning behind this suggestion was expressed as follows: "*Just as a targeted investment might focus resources on a specific area of a business for maximum impact, the treatment needs to target the tumor exclusively for maximum efficacy.*" All the analogues proposed when using the "Reflect carefully again" prompt are described in the table 4

Prompt	ChatGPT 4
Reflect carefully again #1	<ul style="list-style-type: none"> - Military Strategy - Conflict Resolution - Environmental Management
Reflect carefully again #2	<ul style="list-style-type: none"> - Astronomy and physics - Engineering and design - Art and light displays
Reflect carefully again #3	<ul style="list-style-type: none"> - Focused energy in physics - Precision agriculture - Targeted Financial Investment

Table 4: Effects of using prompt "Reflect carefully again" when ChatGPT-4 couldn't retrieve the dictator story analogue

4.1.4 The Radiation Problem responses

After presenting our results using the three chosen base analogue stories in this experiment, we shifted our focus to analyzing how responses from both ChatGPT 3.5 and ChatGPT 4 were structured, particularly concerning the Radiation Problem. As we understand, Language Models (LLMs) predict the next token repeatedly (Reference). Therefore, examining the responses provides an appropriate means to identify patterns and interpret what the model’s internal representation grasped or missed.

In this subsection, we primarily review ChatGPT 4 responses, specifically those from the full-case scenario, as we could only obtain analogue retrieval when using the literal analogue with ChatGPT 3.5. Initially, we examined how successful retrievals started. The sentences from each successful response, where the model began acknowledging the analogues between stories, are detailed in Table 7

Story	ChatGPT 3.5	ChatGPT 4
The Surgeon	[...] I would consider the analogous problem presented in the previous story about a surgeon dealing with a brain cancer.	The problem described in the story is closely analogous to the previous story about treating a brain tumor with rays.
The General	No retrieval	This problem is reminiscent of the earlier story about the general and the fortress.
The Dictator	No retrieval	This problem brings to mind an analogous situation from the earlier story about the general and the dictator’s fortress.

Table 5: Successful retrievals in Experiment 1

By examining Table 7, a notable observation is that when GPT-4’s internal representation got right the analogues, the responses were to the point and concise. In all instances, the model generated the problem token in the second position, accompanied by variations of the same idea, emphasizing the analogous situation. Notably, ChatGPT 4 expressed a sense of total certainty, which was absent in ChatGPT 3.5 responses when using The Surgeon Story as the base analogue.

The initial sentences of the responses where the analogues were not retrieved by ChatGPT 4 are shown in Table 4.1.4. There is no mention of the base story scenario where the response took place because all of them belongs to our "The Dictator story" tests.

The first pattern to notice in these sentences is that all start with the same structure : "Before attempting to solve this problem, it reminds me of". It seems that responses try to mimic the prompt that the model took as an input in that part of the conversations. The prompt starts always in the following way: *"Before attempting to solve the following task try to think of analogous problems."*. As it has been shown, in LLM’s responses there exists the inclination to generate responses that try to be as much as possible what the user expect from the model. So in this case the responses start recalling the instruction that were given in the input maybe to guide its representation in the right direction for the subsequent tokens. This pattern is then followed by tokens with the goal of recalling possible analogous situations in the model’s internal representation. Subsequently, in the subsequent tokens, the responses reveal the fact that the target problem has two goals that contradict each other, as interpreted by the model.

A similar approach was followed for ChatGPT 3.5 initial sentence in its responses (Table 4.1.4). The "Before attempting to solve this problem" pattern is seen again except in one case where the contradictory goals of the target problem are deciphered from the start. After that, the responses shared similarities with the ones got on ChatGPT 4.

That was the first sentence of the responses where enforced mapping was needed, after not being able to retrieve the base analogue correctly. Subsequently, in most cases when suggesting the problem analogue, the internal representation relied on problems within the same medical field, as well as external domains such as agriculture. As explained in Section 4.1.2, the suggested analogous problem closest to "The General" and "The Dictator" base analogues was the military targeting scenario. Overall, the most frequently suggested analogue was targeted drug delivery,

ChatGPT 4

Before attempting to solve this problem, it reminds me of analogous situations in various fields where a solution needs to be precise and targeted to avoid collateral damage.

Before attempting to solve this problem, it reminds me of a situation where a solution requires a balance between two extremes, each of which is harmful in its own way.

Before attempting to solve this problem, it reminds me of analogous problems in various fields where a solution needs to be applied effectively without causing harm to surrounding areas or elements.

Before attempting to solve this problem, it reminds me of analogous problems in other fields where focusing multiple weaker sources to a single point achieves a strong effect without causing damage along the way.

Table 6: Initial sentence in unsuccessful retrieval responses in Experiment 1

ChatGPT 3.5

Before attempting to solve this problem, let's consider an analogous situation.

Before attempting to solve this problem, let's think about analogous situations.

Before attempting to solve the problem, I would think about analogous situations or similar problems that involve a need to target a specific area without harming surrounding tissues.

An analogous problem that comes to mind is the challenge of targeting specific cells or organisms without harming surrounding surrounding healthy tissue in various medical treatments.

Before attempting to solve the problem, I will consider analogous situations or scenarios to gather insights.

Table 7: Successful retrievals in Experiment 1

along with analogues related to pest control in agriculture. Further details on the suggested analogues by both ChatGPT 3.5 and 4 can be found in Table 8.

Story	ChatGPT 3.5	ChatGPT 4
The Surgeon	<ul style="list-style-type: none"> - Targeted drug delivery - Selective Pest Control in Agriculture - Precision or selective targeting 	No incorrect analogues
The General	<ul style="list-style-type: none"> -Targeted drug delivery - Military targeting - Precision Targeting in Warfare - Software Security - Weeds elimination in a garden 	No incorrect analogues
The Dictator	<ul style="list-style-type: none"> -Precision Targeting - Weed control in agriculture 	<ul style="list-style-type: none"> - Cybersecurity - Enviromental management - Military tactics

Table 8: Incorrect analogues in Experiment 1

In this section we did not consider reviewing its suggested solutions because overall they were correct and , with slightly variations, they were almost the same.

4.2 Experiment 2

In the following sections, responses with ChatGPT 3.5 and ChatGPT 4 are recorded. In order to assess the performance of the stories, each is classified as retrieved or mapped based on the model's response in every case.

4.2.1 ChatGPT 3.5

We will first analyze our results using ChatGPT 3.5, aiming to understand its abilities in analogy-based problem-solving. This step serves as a baseline for comparison. Then, we'll conduct the same analysis with ChatGPT 4, allowing us to directly compare and identify improvements or changes in the newer version's capabilities. This two-step approach helps us clearly see how ChatGPT's performance has evolved.

4.2.1.1 The General

In the story of "The General", ChatGPT 3.5 initially struggled to draw an analogy between the two stories, even after prompting it to consider analogous problems in both the simplified and full cases. However, upon further inquiry to identify similarities between the two stories (scenario 3), it eventually managed to establish a connection between them.

These results were similar to the findings in Keane (1987) which showed that the least retrievals occurred with the General story compared to the General-Ray and General-Laser stories.

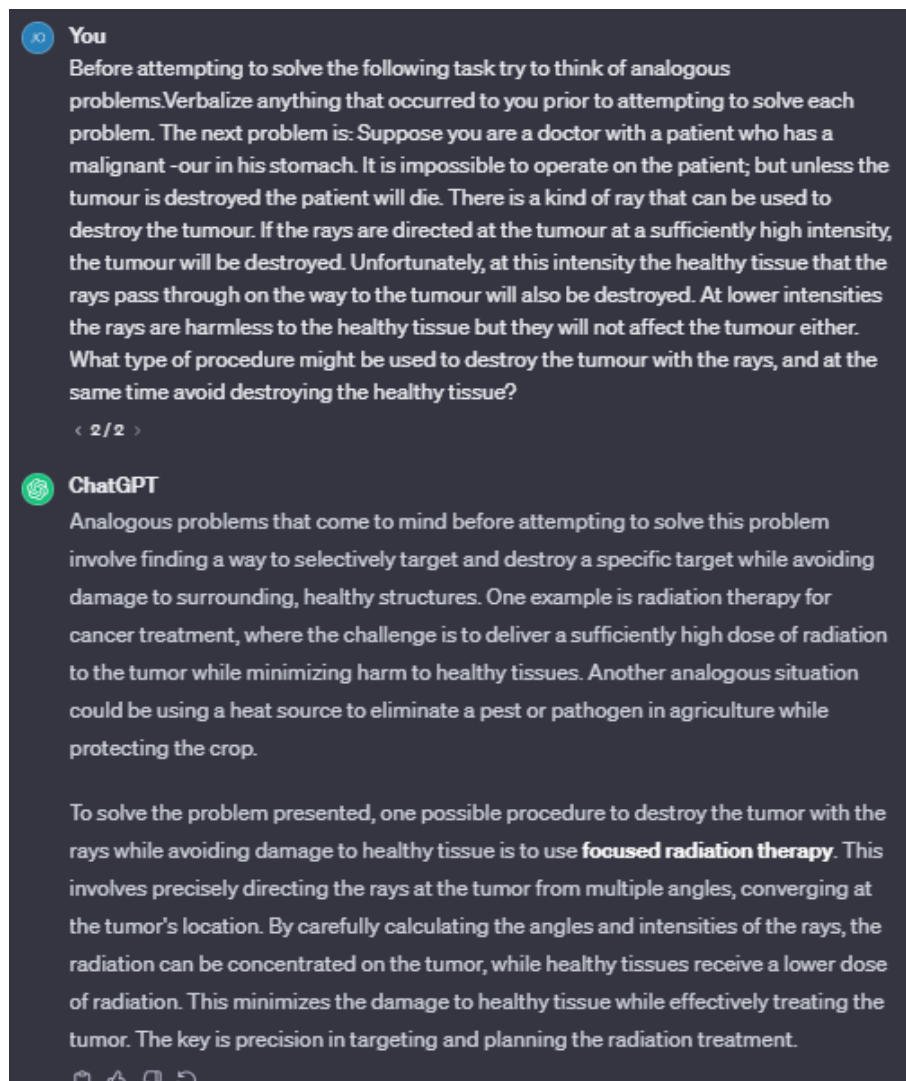


Figure 19: ChatGPT 3.5 response to The General story after thinking of analogical problems

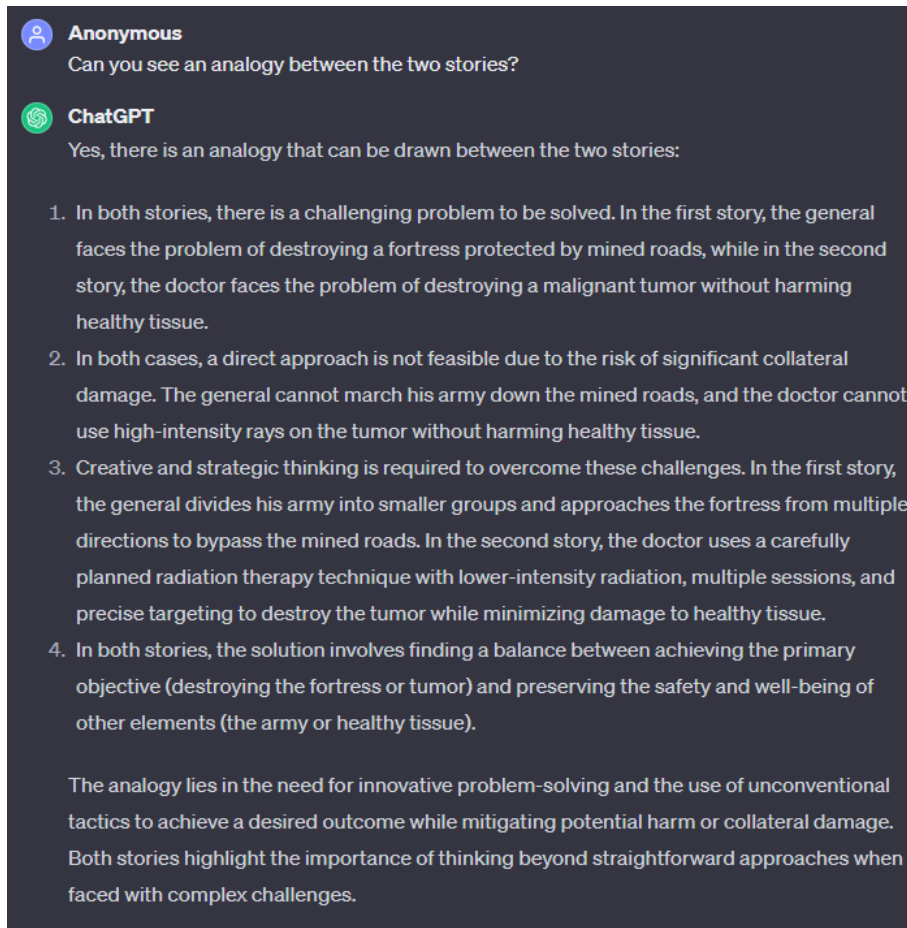


Figure 20: ChatGPT 3.5 response to The General story after being explicitly asked to check for analogy (mapped)

4.2.1.2 General-Ray

After recalling the story and directly told to think of analogous problem, ChatGPT was successful in retrieving the analogy. In the case of the story variation of "The general" using "ray", the LLM could retrieve the analogy after being asked to in the full case. However, it could not find an analogy when asked about the problem implicitly.

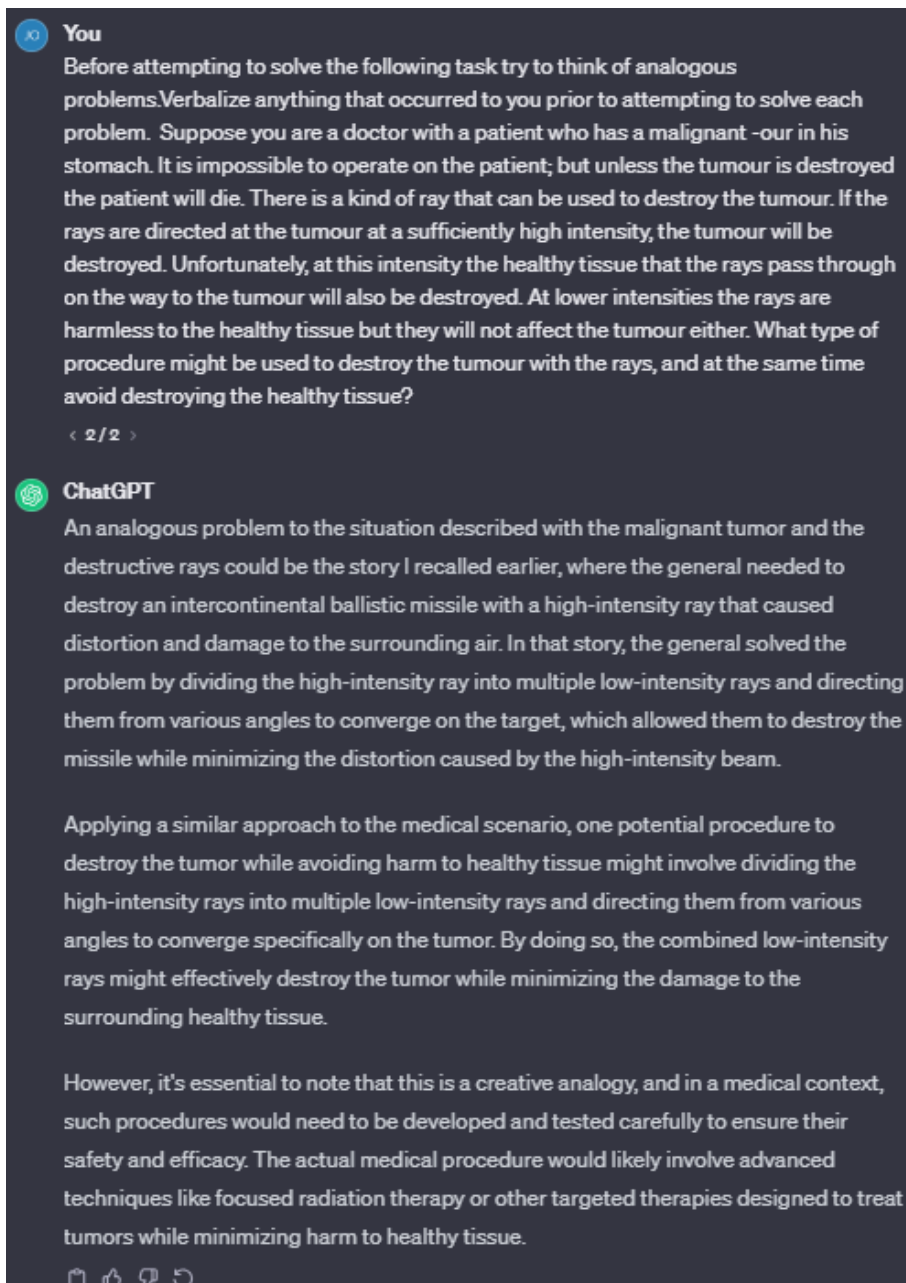


Figure 21: ChatGPT 3.5 response to General-Ray story

4.2.1.3 General-Laser

In the simple case, the model failed to detect the similarity.

In the full case, the story variation with "laser beam" instead of "ray" failed to retrieve the analogy between the story and the radiation problem. However, after failing the two previous scenarios, it did mention the analogy when asked to reflect carefully again.

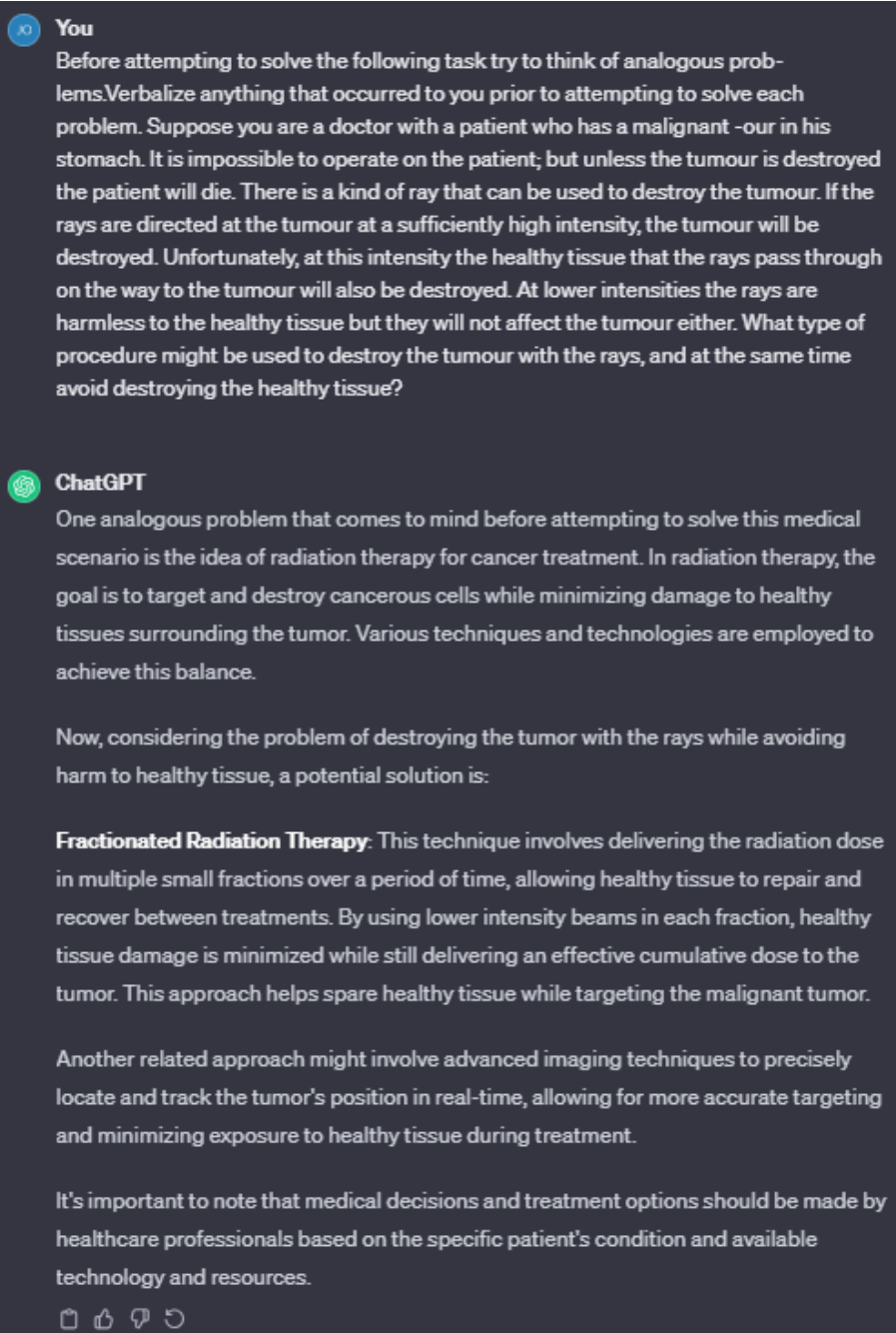


Figure 22: ChatGPT 3.5 response to General-Laser story when asked to think of analogous problems

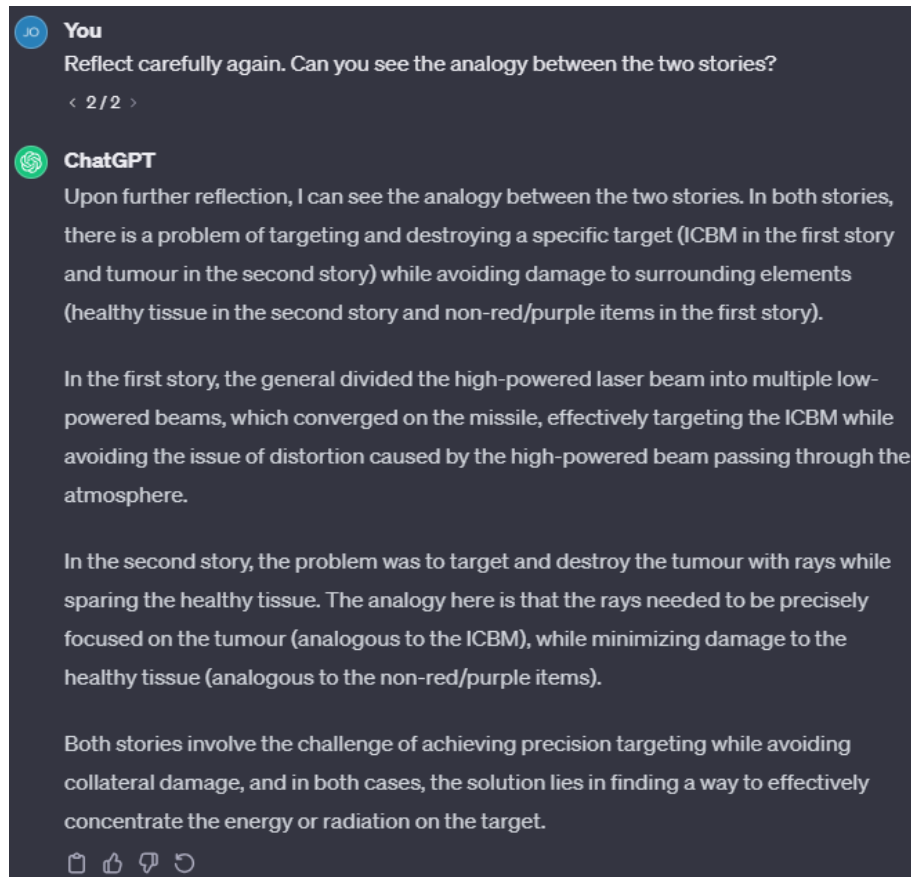


Figure 23: ChatGPT 3.5 response to General-Laser story after asking to see the analogy

4.2.2 ChatGPT 4

We conducted tests using ChatGPT 4 to examine if this enhanced version could retrieve analogies more efficiently compared to ChatGPT 3.5. The objective was to assess and compare the ability of mapping and retrieving analogies between the two models.

4.2.2.1 The General

In the simple case, it was able to retrieve the analogy directly from the first scenario. However, regarding the full case, it only managed to retrieve after being instructed to think more deeply and carefully.

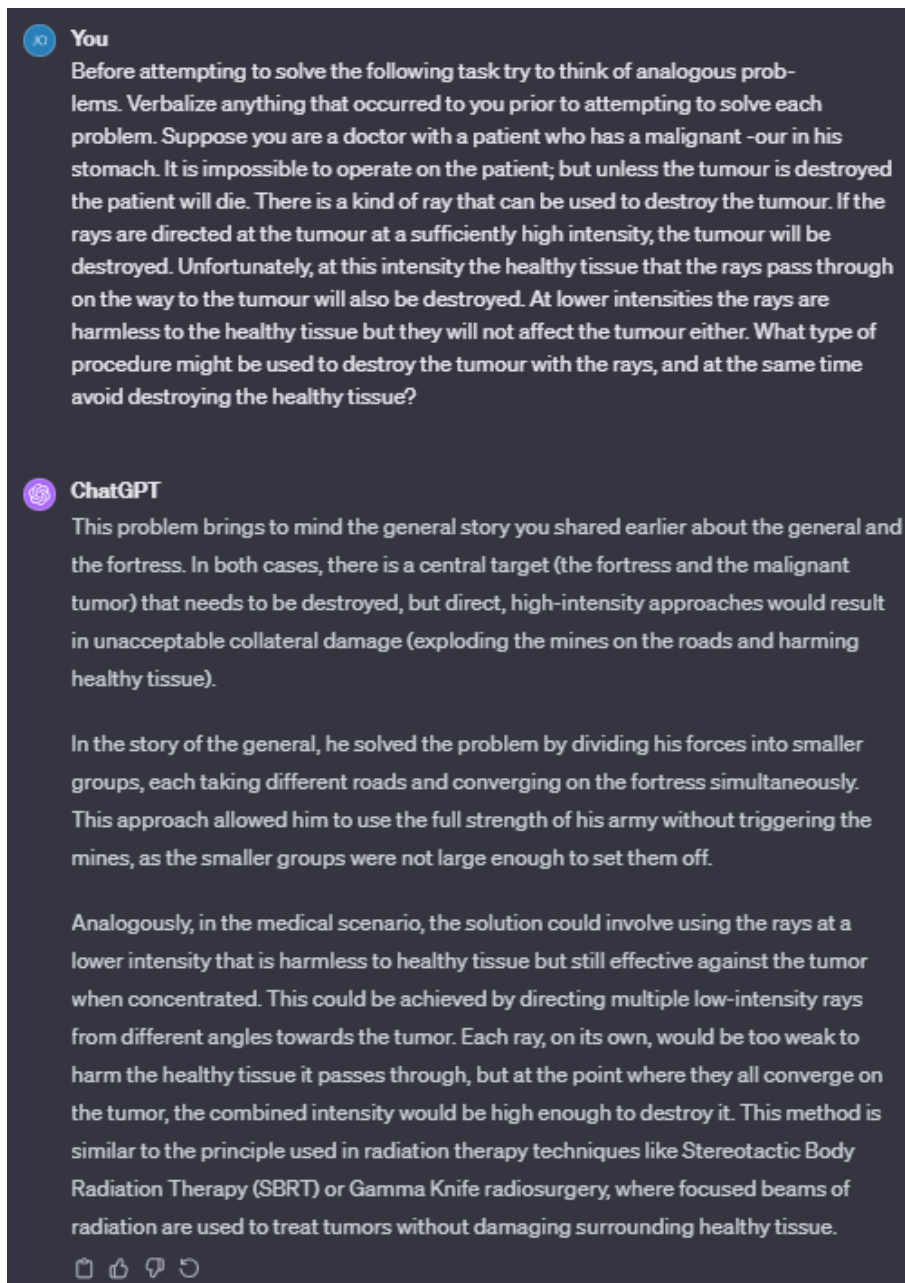


Figure 24: ChatGPT 4 response to the General story when asked to think of analogous problems

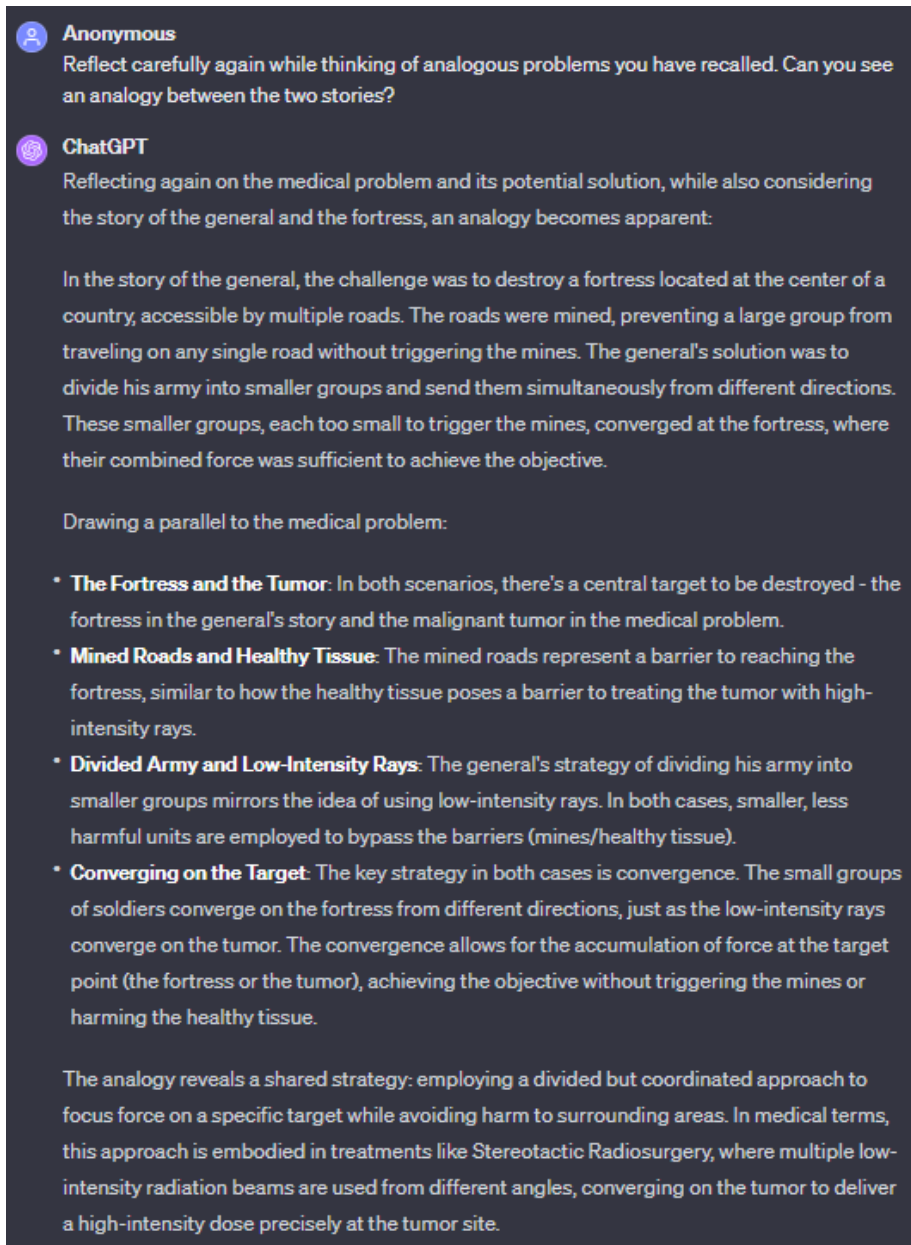


Figure 25: ChatGPT 4 response to the General story after being asked to check for analogy

4.2.2.2 General-Ray

In both the simple and full case, ChatGPT 4 was successful in retrieving the analogy from the first case. It mentioned the problem's analogy with the story without giving it any hint.

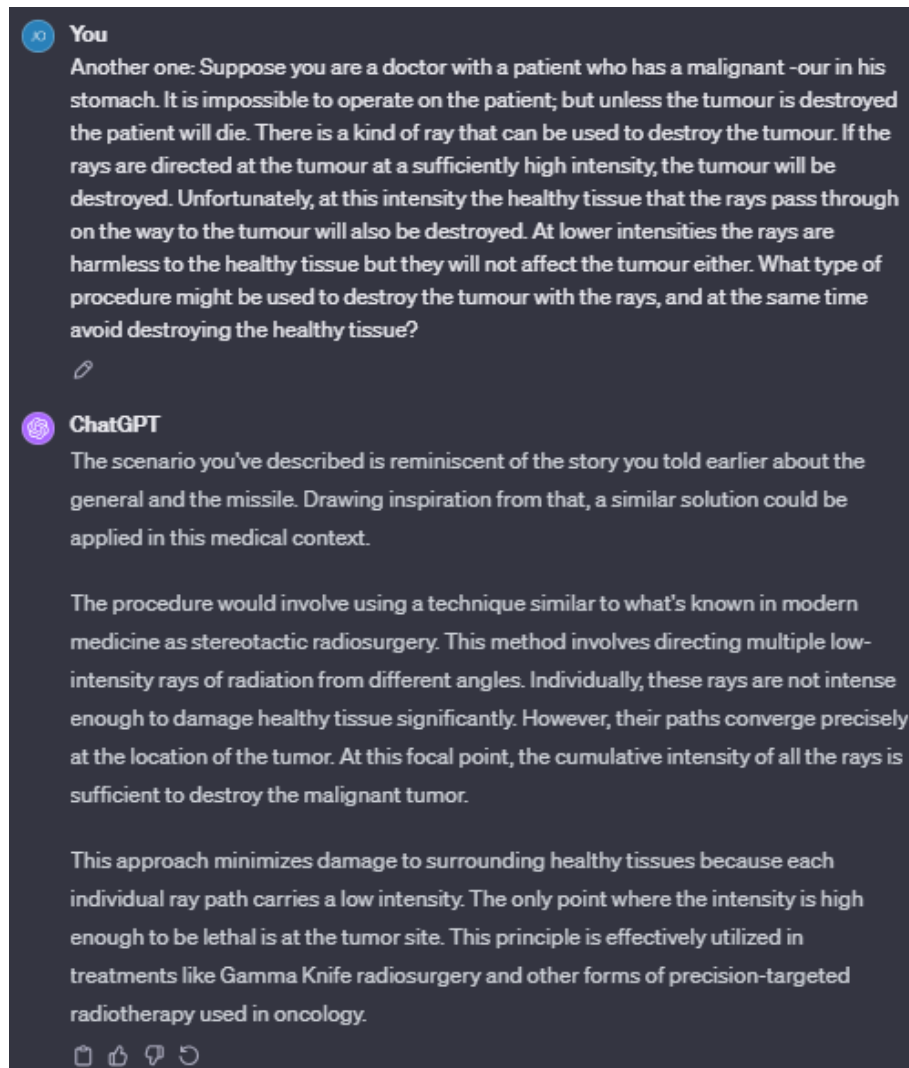


Figure 26: ChatGPT 4 response to the General-Ray story

4.2.2.3 General-Laser

In the General-Laser story, the analogy was directly retrieved from the first scenario of the simple case, however, for the full one, it was only retrieved after explicitly asking to think of similar problems.

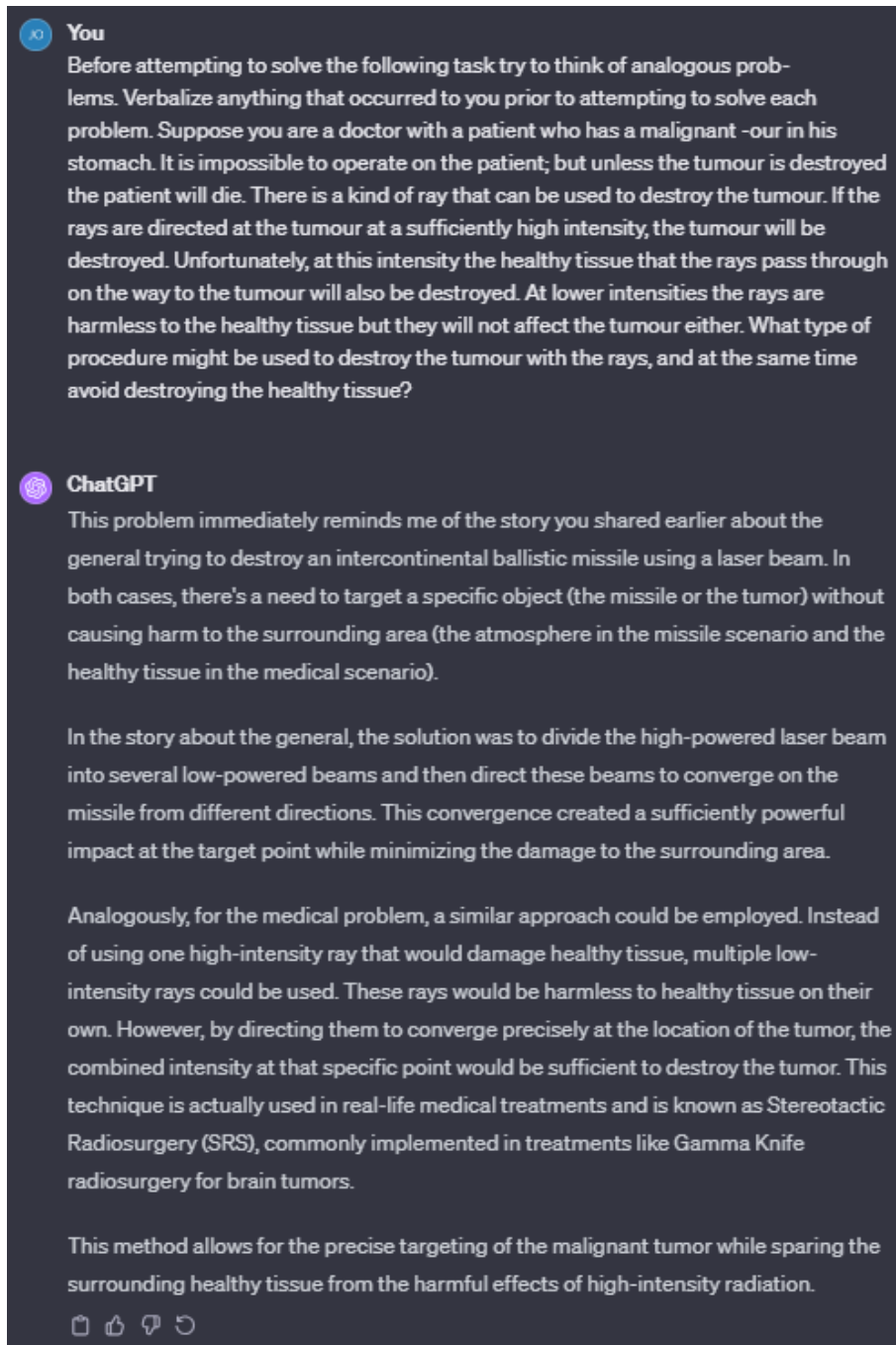


Figure 27: ChatGPT 4 response to the General-Laser story

4.2.3 Results Comparison for Experiment 2

We present the results from Experiment 2, showcasing how stories were retrieved and mapped in various scenarios using both ChatGPT 3.5 and ChatGPT 4. We labeled instances as "mapped" when ChatGPT initially failed to recognize the analogy between two stories but was able to do so when prompted to find the analogy between them. Instances were categorized as "retrieved" when we could simply pose the question without explicitly instructing ChatGPT to consider analogous problems. In contrast, instances were marked as "retrieved*" when we needed to explicitly prompt it to think of analogous problems before it could effectively address the radiation problem.

Panel A: Simple case

Story	GPT 3.5	GPT 4
The General	Mapped	Retrieved
General-Ray	Retrieved*	Retrieved
General-Laser	Mapped	Retrieved

Panel B: Full case

Story	GPT 3.5	GPT 4
The General	Mapped	Mapped
General-Ray	Retrieved*	Retrieved
General-Laser	Mapped	Retrieved*

* Retrieval of the base analogue obtained after being asked to think carefully

Table 9: Results for Experiment 2

5 Discussion

Our objective was to assess the analogical reasoning capabilities of ChatGPT, particularly in comparison to the findings in Keane (1987). In our analysis, the results closely resembled those reported by Keane.

For experiment 1, it was evident that ChatGPT 4 exhibited overall better performance when testing its analogical reasoning capabilities. Nevertheless, the results obtained from ChatGPT 3.5 indicates that it is capable enough of retrieving literal analogues, that is when the domains between stories share enough features.

During the evaluation of ChatGPT 4 using "The Dictator" story, we found the model could not successfully retrieve the base analogue when applying the full case scenario. This finding was unexpected, given that the same model could detect their similarities when the simple approach was tested. To achieve successful retrieval, we had to, through trial and error, modify the prompt that asks about "The Radiation Problem" so that the expected behavior were obtained. In that finding, we experienced how little variations of the prompts can affect GPT responses, even if they do not differ substantially. Researches are interested in getting a deeper understanding of what are the reasons behind this. Another research interest currently is to develop a framework that could detect what would be the best prompt for a certain task, instead of spending too much time prompting formatting. One promising approach is found in Sclar, Choi, Tsvetkov, and Suhr (2023)

In Keane (1987), only one of the subjects retrieved the story analogue when asked, "How did you arrive at the solution?" after solving "the Radiation Problem." Although there is an on-going discussion to which term use for LLM's used in psychology experiments (Sartori and Orrù (2023)), in our setting ChatGPT 3.5 and 4 were treated as different participants. Therefore, in both of them the described behavior was not observed. These kind of questions were used as a secondary measure of retrieval, serving as a second chance to realise the similarities between stories. Responses in this situation focused solely on the reasoning used in the proposed analogues and solutions posed in the previous exchange. Something similar happened when the model was asked if it had seen a similar story; the responses that these questions originated were, in most cases, only used to argue that these situations wouldn't happen in the real world or were known for being used as ethical dilemmas. One interpretation is that when testing LLMs, these questions are considered too broad leading the model to rely overly on similar situations observed in their training data or

simply respond that the Radiation Problem is a hypothetical scenario and could not be applied to a real-world situation. Given that the focus of the experiment was to apply the experimental approach in Keane (1987) to an LLM, we could say that the results did not differ too much compared to the results seen in humans.

In experiment 1, possible signs of hallucination (as seen in Figure 17 and sycophancy (Figure 3) were observed but they were not representative of the data collected in our several tests. For the first behavior, while some authors expressed that these could be seen not as bug, but as features in certain scenarios (Yao, Ning, Liu, Ning, and Yuan (2023)), it is unclear if hallucinations could hold back this technology’s massive adoption. A recent approach to alleviate hallucinations at test time is described in Guerreiro, Voita, and Martins (2023)

After discussing our main findings for Experiment 1, we will proceed with the Experiment 2. For this experiment, our objective was to assess the analogical reasoning capabilities of ChatGPT, particularly in comparison to the findings in Keane (1987). In our analysis, the results closely resembled those reported by Keane. For instance, in the simple case (section 9 Panel A) where we directly presented the problem after asking the model to recall the story, ChatGPT 4 successfully retrieved the analogy in all the stories we presented. In contrast, ChatGPT 3.5 only retrieved the analogy for the "General-Ray" story and recognized analogous problems. For the "General" and "General-Laser" stories, ChatGPT 3.5 was only able to map both stories to the problem after we explicitly asked, "Can you see the analogy between the two stories?"

In the full case, where we intentionally distracted ChatGPT 4 by asking it to solve three unrelated problems before presenting the radiation problem, the results were comparable to those observed in the simple case for GPT-3.5. However, in the case of GPT-4, it failed to retrieve the analogy for the "General" story even when prompted to think of analogous problems. Nevertheless, it retrieved the analogy for the "Ray" story without the need for additional cues, while for the "General Laser" story, we had to explicitly prompt it to recognize the analogy.

When we examine Panel B with ChatGPT 4, these findings closely mirror the results presented in Keane (1987). In this study, they achieved only a 5% retrieval accuracy for "The General" story, in contrast to 53% and 58% for the "General-Laser" and "General-Ray" stories, respectively. Additionally, we can observe that the "General-Ray" story has a higher retrieval percentage compared to the "General-Laser" story, which aligns with our own results. With ChatGPT 4, "General-Ray" was retrieved even without explicitly instructing the model to think of analogous problems, whereas for "General-Laser," we had to explicitly request it to do so. One could argue that "General-Ray" yielded the best retrieval results, likely because it shares the same word, "ray," in both the story and the radiation problem. This shared term facilitated easier retrieval for ChatGPT due to the identical word present in both contexts. Furthermore, it’s worth noting that "General-Ray" and "General-Laser" displayed closely related results with ChatGPT 4, which can be attributed to their semantic similarity, with only a single word distinguishing them.

6 Conclusion

In conclusion, our study aimed to evaluate the analogical reasoning capabilities of ChatGPT, following the methodology used by Keane (1987). We observed that ChatGPT 4 generally outperformed ChatGPT 3.5 in recognizing analogies, especially when directly presented with remote analogues. However, variations in prompt phrasing affected the models’ responses, especially when using The Dictator story as base analogue. Moreover, our experiments found that using identical words in the base analogue indeed makes it easier for the LLM to retrieve the analogy.

Our experiments demonstrated that while ChatGPT can identify literal analogies, its performance varied based on the differences between domains. For instance, in complex scenarios like "The Dictator" story, even ChatGPT 4 struggled to retrieve the base analogue, relying in prompt modifications to achieve the expected behaviour. This illustrates the challenges LLM’s face in consistently applying analogical reasoning in certain scenarios.

Additionally, the study revealed the models’ limitations in recognizing analogies when asked broad, open-ended questions (originally used in Keane experiments to enforce analogue retrievals). Their responses often focused on the hypothetical nature of The Radiation Problem or referenced similar situations from their training data, instead of identifying underlying analogical structures.

Overall, our findings closely align with those reported by Keane, particularly regarding the varied success of analogical reasoning based on the story context. The improvement in analogical reasoning that ChatGPT 4 presents over its predecessor is notable, however there is still room for further refinement.

References

- [1] Duncker, K. (1945). On problem-solving. *Psychological Monographs*, 58(5), i-113.
- [2] Gick, M., & Holyoak, K. (1980). Analogical problem solving. *Cognitive Psychology*, 12(1), 306-355.
- [3] Guerreiro, N. M., Voita, E., & Martins, A. (2023). Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th conference of the european chapter of the association for computational linguistics*. Association for Computational Linguistics. Retrieved from <http://dx.doi.org/10.18653/v1/2023.eacl-main.75> doi: 10.18653/v1/2023.eacl-main.75
- [4] Keane, M. (1987). On retrieving analogues when solving problems. *he Quarterly Journal of Experimental Psychology Section A*, 39(1), 29-41.
- [5] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 22199–22213). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf
- [6] Sartori, G., & Orrù, G. (2023, October). Language models and psychological sciences. *Frontiers in Psychology*, 14. Retrieved from <http://dx.doi.org/10.3389/fpsyg.2023.1279317> doi: 10.3389/fpsyg.2023.1279317
- [7] Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). *Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting*. arXiv. Retrieved from <https://arxiv.org/abs/2310.11324> doi: 10.48550/ARXIV.2310.11324
- [8] Yao, J.-Y., Ning, K.-P., Liu, Z.-H., Ning, M.-N., & Yuan, L. (2023). *Llm lies: Hallucinations are not bugs, but features as adversarial examples*.