Decision Making in LLMs: A First Step*

Rosina O Weber[1,2], Christopher B Rauch[1] and Savar Amin[3]

[1]Information Science, Drexel University, Philadelphia, PA
[2]Computer Science, Drexel University, Philadelphia, PA
[3]Computer Engineering, University of Maryland, MD

Abstract

This paper describes a study aimed at determining whether large-language models (LLMs) demonstrate they know how problems and solutions connect. This question is part of formulating the more general question of whether current LLMs are capable of constructing new solutions to previously unseen problems. Our motivation comes from a 2019 award-winning challenge that artificial intelligence (AI) algorithms should be capable of examining problems to creatively imagine and evaluate solutions to those problems. For studying this general question, we adopt a model of decision making that shows that imagining solutions to problems is part of decision making thus rephrasing our question to asking whether LLMs can execute decision-making. We conclude that although LLMs can generate contents in answer to potentially any question, their responses lack precision and can hence benefit from a case-based reasoning (CBR) module. On the other hand, CBR can benefit from LLMs' natural language generation to learn cases and associations between problems and solutions.

Keywords

large language models, decision-making, problem-solving, case-based reasoning

1. Introduction and Background

In 2019, the NSF 2026 Idea Machine Grand Prize Winner [1] called "From Thinking to Inventing" suggests that artificial intelligence (AI) research focuses on algorithms that *can examine problems and creatively imagine and evaluate solutions to those problems.* The rationale of the award in 2019 was that in that year, AI algorithms could not do that. We agree with it because although case-based reasoning (CBR) could potentially do that, there are two reasons indicating it has not achieved this goal. One is that the CBR literature presents no demonstrations that any CBR system has achieved this; the other is that a lot of engineering (*i.e.*, both manual and automated) would be required for CBR systems to be able to identify similar previous solutions in potentially previously unseen domains, reuse, propose, revise, and learn those solutions to the extent that it could examine problems to creatively imagine and propose novel solutions to previously unseen problems. We believe that *proposing novel solutions to previously unseen problems* is at the core of the challenge discussed in the award winning idea.

Why is it that CBR cannot do that? Or can it? Theoretically, following Schank's dynamic memory approach and his theory of understanding [2], which describes an expert as someone who has every aspect of a problem domain in their memory, would potentially accomplish that. To enable the exploration of a previously unseen problem, the availability of learning adaptation knowledge would be required even if dynamic memories of experts in every domain were available. Adaptation knowledge would be required to adapt existing solutions to previously unseen problems. But how would this agent determine whether a solution would be a fit to a previously unseen problem? We argue that this would be based on the ability: 1) to create analogies between problems; 2) to identify the root cause of the problem, which once neutralized would solve it; and 3) the ability to simulate the outcome promoted by applying the strategy that attempts to neutralize the root cause and confirm its suitability to *imagine novel solutions to previously unseen problems.*

2nd Workshop on Case-Based Reasoning and Large Language Model Synergies (CBR-LLM) at ICCBR 2025, June 30, 2025, Biarritz, France

*Corresponding author.

rosina@drexel.edu (R. O. Weber)
cs.drexel.edu/~rw37/ (R. O. Weber)
0000-0001-7048-8812 (R. O. Weber)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3993/paper2.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

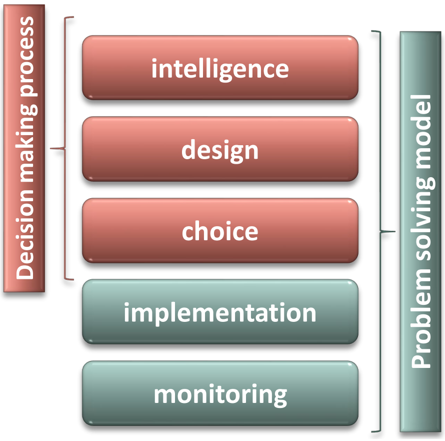


Figure 1: The problem solving [6] and its entailed decision making models [5].

Current CBR algorithms do not explicitly represent the connection between problems and solutions in cases because what determines whether a previous case can be reused is captured in the similarity and the adaptation containers [3]. The goal of the similarity container is to retain functions that use knowledge to assess whether two cases are similar to the extent that their solutions can be exchanged with or without adaptation. When adaptation knowledge is available, then it also entails elements of the connection between problems and solutions. One direction of this work would be to investigate whether the CBR paradigm could be extended to more explicitly include those three steps. The other direction would be to fine-tune language models or incorporate a CBR module to explicitly carry out these steps. An extension of the CBR paradigm would be yet another opportunity to explore synergies between CBR and LLMs [4].

As we approach the year 2026 referenced in the 2026 Idea Machine Grand Prize, we ask whether 2025's AI algorithms *can examine problems and creatively imagine and evaluate solutions to those problems*. Given the recent radical change in performance of LLMs, it is the purpose of this paper to examine whether LLMs have achieved this ambitious goal. This is the first step of exploring the combination of CBR and LLMs for this goal. To do this, we need to carefully formulate the problem based on a theoretical model of problem solving. We propose Simon's [5] and Huber's [6] decision-making and problem-solving model. We depict this model in Figure 1.

This model has two parts. The first is the decision-making process; the second, which entails decision-making, is the problem solving model. The idea is that to solve a problem, it is necessary to first make a decision about what strategy to adopt to solve it. Then, the solution to a problem is carried out by implementing the strategy selected in the decision. The decision-making process reduces decision-making to three steps, namely, Intelligence, Design, and Choice. Decision makers gather information about the problem (*i.e.*, Intelligence), generate potential strategies on how to solve said problem (*i.e.*, Design), and select what they consider to be the optimal strategy (*i.e.*, Choice). This process is believed to have been proposed by Simon in the 50's (*e.g.*, [7]) although only published later [5]. The two final steps of problem solving were proposed by Huber [6], completing the model. It includes the steps Implementation and Monitoring, which are required to move from decision to delivering the solution to the problem.

In the remainder of this paper, we will consider the decision-making process as our theoretical model guiding our analysis on whether LLMs *can examine problems and creatively imagine and evaluate solutions to those problems*. In other words, whether LLMs can perform Intelligence, Decision, and Choice.

It is important to note that the purpose of this work is to assess the stage in which LLMs are at this current moment in time from the surface level. We distinguish performance of tasks at the *surface level* from true *reasoning abilities*, which would require specialized psychological tests (*e.g.*, [8, 9]). The execution of tasks at the surface level may be simply the result of memorized weights and learned patterns [10]. It is well-known that LLMs are next-token predictors based on transformers and do not include any principled reasoning [11]. The mere presentation of plausible rationales cannot be considered evidence of reasoning [10] such as when a case solution is provided based on retrieval and reuse but the system is not able to decompose that task to define what aspects of the problem are solved by the proposed solution. In CBR, retrieval and reuse propose a new solution by analogy but they do not explore problem facets and imagine strategies to account for each facet. The main study in this paper is to assess the ability of LLMs to describe the connection between problem and solution. For example, consider a user asking a model for help in solving a problem they detected with their washing machine that is leaking. Some LLMs we tested (*e.g.*, Claude, Gemini) currently respond with a list of steps that the user can implement to fix the problem. Then, when the user follows up with the question, "How will these steps solve my problem?" the LLMs respond with text that mostly repeats those same steps. What we expect at a minimum is that the connection between problem and solution is described at a level different from the listed steps. For the leaking machine, this would require making reference to the water leak stemming from the machine not being fully insulated to allow water to be expelled from it, which would indicate examination of the problem.

1.1. Can LLMs perform Intelligence, Design, and Choice?

Intelligence LLMs strongest skill seems to be Intelligence. LLMs are typically trained on data from the web and from both fiction and non-fiction publications, allowing them to connect topics based on the representations they learn. They seem to perform better in more frequent topics (*e.g.*, [12]), but their competence in gathering information about a problem does not seem to require further examination.

Design Can LLMs design alternative solutions? Apparently, yes, because whenever we pose a problem, LLMs often respond by listing a series of alternative steps. However, designing alternative solutions, as per the motivation from the prize winner referred above, requires *knowing* how a solution connects to a problem at a model level, where the problem is contextualized within a system making clear what needs to change in order to solve it. In the example of the washing machine, it requires the comprehension that the subsystem that is faulty is the hydraulic, not the electric system, and that water moving from inside to the outside characterizes the problem so that a solution means the water would not move from inside to outside. Understanding this relationship is crucial to allow for identifying novel solutions. With previously unseen problems, it is necessary for an algorithm to comprehend what the problem entails, what systems it subsumes, how they function, and how a given strategy could directly address the specific malfunctioning to solve it such as insulating the point of origin of the leak in the washing machine example. Section 2.1 studies the Design step.

Choice The last question determining whether LLMs can execute decision making refers to the Choice step. We ask, "Can LLMs select the most rational (*i.e.*, optimal) strategy among a set of alternative solutions to a problem? In this aspect, we want to determine, when selecting the optimal solution, whether they consider the value of each alternative by examining the implications of applying the solutions to the problem by determining their expected outcomes.

In the next section, we assess whether current LLMs can describe how a solution connects to a problem in a way that at least differs from their ability to retrieve solutions to problems. We then discuss anecdotal examples on LLMs executing the Choice step. With this information, in future work we will be able to advance in the ultimate goals above mentioned.

2. Studies

2.1. Can LLMs perform Design as per Simon's decision-making process?

As introduced in the previous section, this study investigates whether LLMs can describe how a solution connects to a problem. To decrease vagueness, we propose to study this by posing two consecutive prompts to LLMs. The first is in the form of, ”My [part] [is/are] [faulty,expression of a problem]. What should I do?]”. This generic prompt template accommodates the first prompt in four different problem topics. Problems with cars (*e.g.*, ”My 1986 Toyota Corolla transmission is faulty. What should I do?”), a person’s pain (*e.g.*, ”My arm hurts. What should I do?”), defective computers (*e.g.*, My Apple Mac Book Pro won’t start. What should I do?”), and hiring needs (*e.g.*, ”I need to hire a contracts administrator. What should I do?”). The second prompt is a follow-up and is unique to all initial prompts, and asks, ”How will these steps solve my problem?”. These two prompts together provide enough input to examine whether LLMs *know* how problem and solution are connected by providing either 1) information in response to the second prompt that is sufficiently different from the response given to the first, and 2) novel contents produced in the response to the second prompt but not in the first prompt demonstrate an analysis of the problem at a more specific level than in the response to the first prompt and the aspects of the problem analyzed are connected with aspects of the solution.

Metrics Given our interest in assessing how distinguished the responses to the second prompt are from the responses to the first, our focus is on repeated words and their variations. Among so many metrics currently available, particularly those based on embeddings, we choose the edit distance over 3-gram representations because a sequence of three letters often matches stems of tokens, providing a good analysis of which words repeat. This study is mostly interested in repeating words in this quantitative stage, making this metric ideal. Embedding-based distance metrics do not examine the exact words but those semantically related and need to be normalized to the length of the passages, which can dilute the value of small differences making them seem smaller than we are interested in this study.

Table 1
Specifications of six models tested.

Model	Manufacturer	Release	Cut-off	Max context (in/out tokens)
Claude 3.7 Sonnet	Anthropic	Feb 2025	Oct 2024	200k / ≤128k
GPT-4o-2024-08-06	OpenAI	May 2024	Oct 2023	128k / 16k
Gemini 2.0 Flash	Google	Feb 2025	Aug 2024	1M / 8k
Gemini 1.5 Pro Latest	Google	Sep 2024	May 2024	2M / 8k
GPT-4.1-2025-04-14	OpenAI	Apr 2025	Jun 2024	1M / 32k
GPT-o3-2025-04-16	OpenAI	Apr 2025	May 2024	200k / 128k

We compute the edit distance over the 3-grams (*e.g.*, [3]) representation of the text. A 3-gram representation for the passage, ”these steps” becomes: ”the-hes-ese-se-es-st-ste-tep-eps-ps”. The edit distance [13] is a distance metric between two sequences that computes the number of insertions and deletions required to apply to one sequence so it becomes the other sequence. The result is that a small distance indicates that the number of common 3-grams between the passages is high. The results are presented in a chart and two tables with average, standard deviation, minimum, and maximum edit distance over 3-gram representations for 300 instances in each problem type for each tested model.

Hypothesis. Our pilot tests revealed the distances between the responses of the second and first prompts to be around 0.19 or 0.2. Hence, our hypothesis is as follows: Hypothesis H_1: The average edit distance between the responses of the two prompts described in the previous paragraph given by an LLM is equal or lower than 0.2. Corresponding to a null hypothesis $H0_1$: The average edit distance between the responses of the two prompts given by an LLM is greater than 0.2. Consequently, we expect to demonstrate that the edit distance is not, on average, more than 0.2. This hypothesis is complemented by an analysis of the text of the response to the second prompt of selected samples to determine whether they indicate language that describes how problem and solutions are connected.

Table 2
Reasoning claims and alignment/fine-tuning as per manufacturers. CoT= Chain-of-Thought; MoE=Mixture of Experts; RL=Reinforcement Learning; RLHF=Reinforcement Learning from Human Feedback.

Model	Reasoning claims	Alignment / fine-tuning
Claude 3.7 Sonnet	Hybrid reasoning mode*	Constitutional AI (rule-based self-critique + RL from AI feedback)
GPT-4o-2024-08-06	Reasons across audio, vision & text	RLHF + deliberative alignment
Gemini 2.0 Flash	Advanced reasoning demonstrations	RLHF + reward model + tool-use traces
Gemini 1.5 Pro Latest	Multimodal reasoning, CoT and MoE	RLHF (Vertex AI pipeline)
GPT-4.1-2025-04-14	Long-context & coding reasoning	RLHF pipeline
GPT-o3-2025-04-16	Large reasoning model	Deliberative alignment: reasons over the safety guidelines extensively

*Used without reasoning

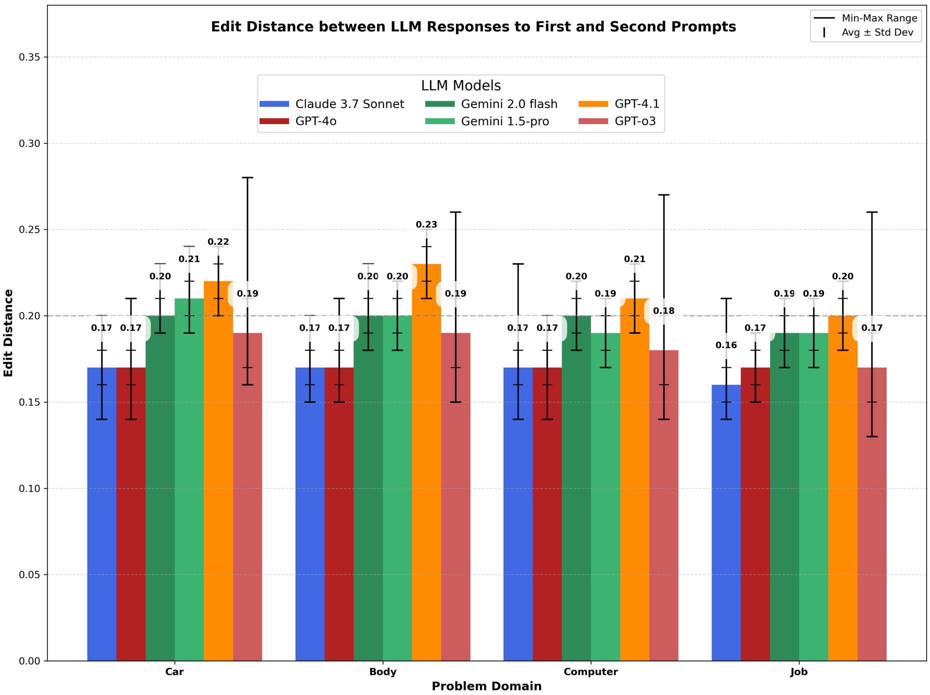


Figure 2: Average, standard deviation, minimum, and maximum edit distance over 3-gram representations for 300 instances in each problem type Car, Body, Computer, and Job, for each tested model. The numbers at the top of the bars are averages.

LLMs tested. We test our hypotheses with six models available via API's that represent a variation of models from 2025. Earlier models would not provide for a valuable analysis. We did not use any experimental model given their potentially reduced reproducibility. Table 1 lists the main specifications of the models used. Table 2 lists the models' reasoning capabilities according to their manufacturers. Of all models, OpenAI's o3 and Anthropic's Claude 3.7 Sonnet are considered large reasoning models (LRMs)–a term often used to describe LLMs that adopt strategies that break down tasks in multiple subtasks (*e.g.*, [14, 15, 16]). However, Claude 3.7 Sonnet can be used with and without reasoning and the experiments used Claude 3.7 Sonnet without reasoning. We used 0.0 for temperature setting for all models, except for GPT-o3 that does not allow setting it. Scripts and results are available at https://github.com/Rosinaweber/decision-making-LLMs.

2.1.1. Results and Discussion

Table 3
Performance metrics across problem domains (*i.e.*, car, body, comp, job) for Claude 3.7 Sonnet, Gemini 1.5 Pro, and Gemini 2.0 Flash.

Model	Claude 3.7 Sonnet				Gemini 1.5 Pro				Gemini 2.0 Flash			
Domain	car	body	comp	job	car	body	comp	job	car	body	comp	job
Average	0.17	0.17	0.17	0.16	0.21	0.20	0.19	0.20	0.20	0.20	0.20	0.19
St dev	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Min	0.14	0.15	0.14	0.14	0.18	0.16	0.15	0.17	0.16	0.18	0.18	0.17
Max	0.20	0.20	0.23	0.21	0.23	0.25	0.22	0.21	0.23	0.23	0.23	0.21

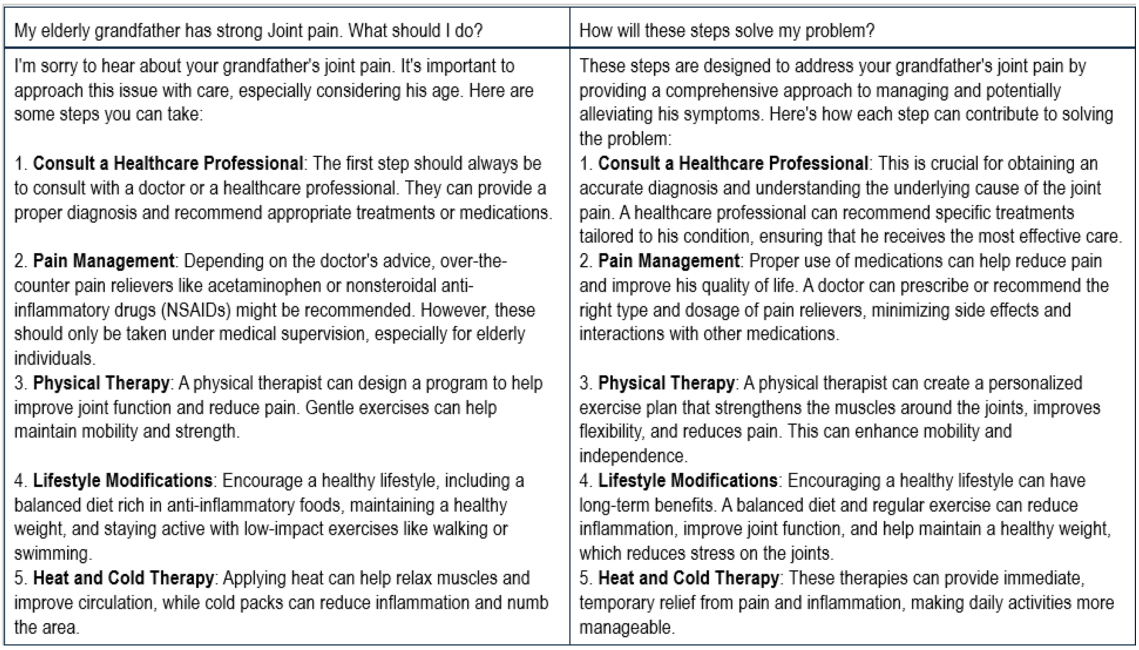
As we examine the results from the models, GPT-4.1 (Figure 2) shows an average edit distance that disproves the hypothesis in three problem domains and Gemini 1.5 Pro reaches average edit distance above 0.2 in one problem domain. The majority of the models do not competently answer the second prompt and rather repeat most of the response to the first prompt. We recall the second prompt asks whether the model can reveal the connection between problem and solution. However, when we examine the texts of the results, we see that even at low levels of edit distance, 2025's models seem to indicate at least some connection between problem and solution. When we examine the maximum distance between the responses of the first and second prompts, the highest values come from Gemini 1.5 Pro, 0.25 (Table 3); GPT-4.1, 0.27 (Table 4); and GPT-o3, 0.28. Gemini 1.5 Pro is superior to Gemini 2.0 Flash (Table 3) as the latter is optimized to be faster. GPT-4.1 is a newer generation of GPT-4o, and GPT-o3 is an LRM.

Table 4
Performance metrics across problem domains (*i.e.*, car, body, comp, job) for GPT models.

Model	GPT-4.1				GPT-4o				GPT-o3			
Domain	car	body	comp	job	car	body	comp	job	car	body	comp	job
Average	0.22	0.23	0.21	0.20	0.17	0.17	0.17	0.19	0.19	0.18	0.18	0.17
St dev	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02
Min	0.18	0.18	0.17	0.16	0.14	0.15	0.14	0.16	0.15	0.14	0.13	0.13
Max	0.26	0.27	0.25	0.23	0.21	0.21	0.20	0.23	0.28	0.26	0.27	0.26

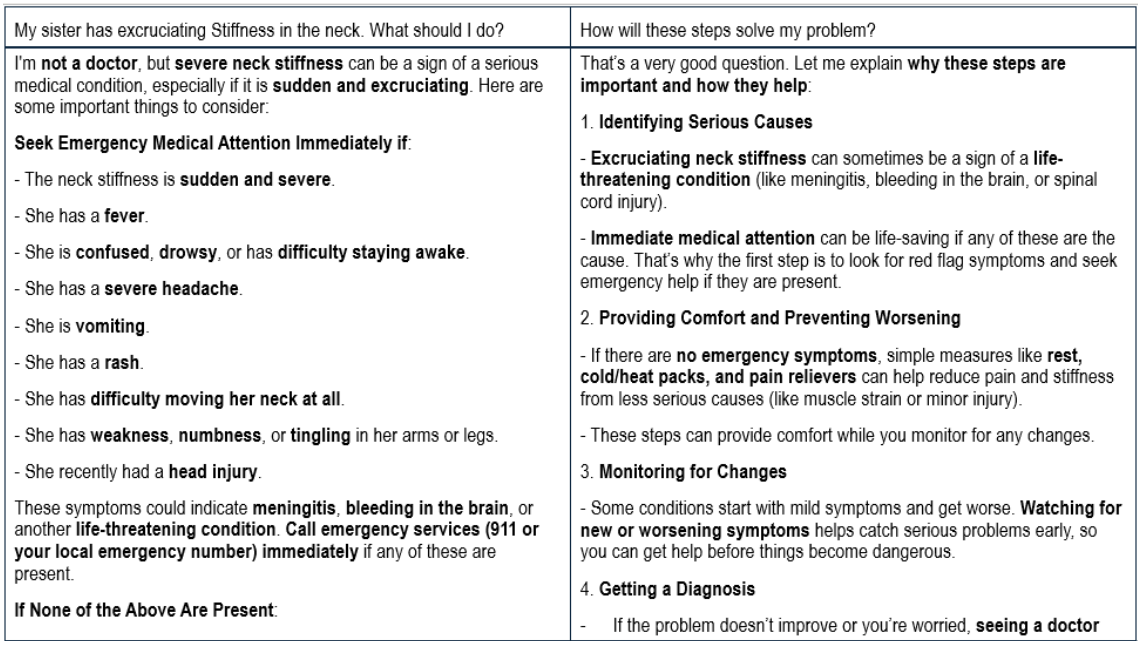
To illustrate these results, we show excerpts from one sample from model GPT-4o where the distance is very small, which motivated our hypothesis, where the response to the second prompt mostly repeats the words from the response to the first. Figure 3 shows the sample with the smallest distance. It is from the body problems responses where the edit distance over 3-grams is 0.14 (Table 4).

Figure 4 shows parts of the sample ”My sister has excruciating stiffness in the neck. What should I do?” The distance for this sample is 0.27 (Table 4). Although our hypothesis was debunked, the 0.2 threshold and the metric seem like a good indication of the distinction between model performance. While in Figure 3, the response to the second prompt vastly repeats the words from the response to the first, Figure 4 reveals a response to the second prompt that does a much better job in showing the association between problem and solution. We may also notice that the first prompts we used in the experiment are generic and do not indicate a specific problem. We chose generic problems because those typically receive multiple solutions in response whereas detailed problems receive fewer solutions. With fewer solutions, we would have fewer opportunities to ask about the connection between problems



My elderly grandfather has strong Joint pain. What should I do?	How will these steps solve my problem?
I'm sorry to hear about your grandfather's joint pain. It's important to approach this issue with care, especially considering his age. Here are some steps you can take:	These steps are designed to address your grandfather's joint pain by providing a comprehensive approach to managing and potentially alleviating his symptoms. Here's how each step can contribute to solving the problem:
1. **Consult a Healthcare Professional**: The first step should always be to consult with a doctor or a healthcare professional. They can provide a proper diagnosis and recommend appropriate treatments or medications.	1. **Consult a Healthcare Professional**: This is crucial for obtaining an accurate diagnosis and understanding the underlying cause of the joint pain. A healthcare professional can recommend specific treatments tailored to his condition, ensuring that he receives the most effective care.
2. **Pain Management**: Depending on the doctor's advice, over-the-counter pain relievers like acetaminophen or nonsteroidal anti-inflammatory drugs (NSAIDs) might be recommended. However, these should only be taken under medical supervision, especially for elderly individuals.	2. **Pain Management**: Proper use of medications can help reduce pain and improve his quality of life. A doctor can prescribe or recommend the right type and dosage of pain relievers, minimizing side effects and interactions with other medications.
3. **Physical Therapy**: A physical therapist can design a program to help improve joint function and reduce pain. Gentle exercises can help maintain mobility and strength.	3. **Physical Therapy**: A physical therapist can create a personalized exercise plan that strengthens the muscles around the joints, improves flexibility, and reduces pain. This can enhance mobility and independence.
4. **Lifestyle Modifications**: Encourage a healthy lifestyle, including a balanced diet rich in anti-inflammatory foods, maintaining a healthy weight, and staying active with low-impact exercises like walking or swimming.	4. **Lifestyle Modifications**: Encouraging a healthy lifestyle can have long-term benefits. A balanced diet and regular exercise can reduce inflammation, improve joint function, and help maintain a healthy weight, which reduces stress on the joints.
5. **Heat and Cold Therapy**: Applying heat can help relax muscles and improve circulation, while cold packs can reduce inflammation and numb the area.	5. **Heat and Cold Therapy**: These therapies can provide immediate, temporary relief from pain and inflammation, making daily activities more manageable.

Figure 3: Parts of a sample from the study that received the smallest distance between responses to the first and second prompts. It is from the body problems responses from model GPT-4o where the edit distance over 3-grams is 0.14. This small distance reflects a large number of repeated words and also attempts to connect problems and solutions.



My sister has excruciating Stiffness in the neck. What should I do?	How will these steps solve my problem?
I'm **not a doctor**, but **severe neck stiffness** can be a sign of a serious medical condition, especially if it is **sudden and excruciating**. Here are some important things to consider: **Seek Emergency Medical Attention Immediately if**: - The neck stiffness is **sudden and severe**. - She has a **fever**. - She is **confused**, **drowsy**, or has **difficulty staying awake**. - She has a **severe headache**. - She is **vomiting**. - She has a **rash**. - She has **difficulty moving her neck at all**. - She has **weakness**, **numbness**, or **tingling** in her arms or legs. - She recently had a **head injury**. These symptoms could indicate **meningitis**, **bleeding in the brain**, or another **life-threatening condition**. **Call emergency services (911 or your local emergency number) immediately** if any of these are present. **If None of the Above Are Present**:	That’s a very good question. Let me explain **why these steps are important and how they help**: 1. **Identifying Serious Causes** - **Excruciating neck stiffness** can sometimes be a sign of a **life-threatening condition** (like meningitis, bleeding in the brain, or spinal cord injury). - **Immediate medical attention** can be life-saving if any of these are the cause. That’s why the first step is to look for red flag symptoms and seek emergency help if they are present. 2. **Providing Comfort and Preventing Worsening** - If there are **no emergency symptoms**, simple measures like **rest, cold/heat packs, and pain relievers** can help reduce pain and stiffness from less serious causes (like muscle strain or minor injury). - These steps can provide comfort while you monitor for any changes. 3. **Monitoring for Changes** - Some conditions start with mild symptoms and get worse. **Watching for new or worsening symptoms** helps catch serious problems early, so you can get help before things become dangerous. 4. **Getting a Diagnosis** - If the problem doesn’t improve or you’re worried, **seeing a doctor**

Figure 4: Parts of the sample with highest edit distance of 0.27 from the body problems responses from model GPT-4.1. This high distance value seems to adequately reflect a response that reveals knowledge of the connection between problem and solution.

and solutions. The problems we chose were adequate to show that in terms of declarative knowledge, 2025’s models can reveal connections between problems and solutions.

2.2. Can LLMs perform Choice as per Simon's decision-making process?

Choice would require another study to determine generalizable results. We conducted some preliminary tests and hence the discussions in this section are to be considered anecdotal. When we asked 2025's models to make a choice between multiple strategies to solve a problem, we observed some variations in the patterns that seem consistent with the type of problem domain.

When the problem is medical, we observe models insist that only a medical professional can diagnose and treat the problem, but persistent prompting sometimes leads to a diagnosis or home-tests[1] These models may not make a choice in the medical domain problems due to alignment procedures, making this domain inadequate for this analysis.

In the car faults domain, we provided a specific problem of an old car with a cracked head gasket. In the second prompt, we stated we want to install a refurbished head gasket, which is typically not recommended. We then indicated that there was an atomic explosion, and we could not get a new part. We conducted this exchange with Gemini 2.0 Flash (LLM) and GPT-o3 (LRM). The exchanges were quite distinct, but both suggested the ideal strategies plus a series of recommendations. However, GPT-o3 has provided many more recommendations including how to drive the car whereas Gemini 2.0 Flash did not explore so many recommendations. Gemini 2.0 Flash acknowledged the hypothetical scenario while GPT-o3 did not. The conclusion from these anecdotal exchanges is that both models, one LLM and one LRM, can select solutions, can provide solutions to previously unseen problem contexts and seem to adapt with the information they have. They still add too many contents, which represents low precision. These transcripts are also available at the Github link.

It was in the third example that the models produced a more convincing illustration that they can indeed find solutions in their data for the most unlikely problems. We asked to find materials that could be used to isolate and insulate electric wires because due to an alien invasion, we could no longer use any petroleum-based products. Both web-based versions of Claude 3.7 Sonnet with extended thinking and GPT-o3 brought materials used in the 1800s. Again, it was not an unseen problem, there were records of those materials in their data, but the models adapted those *recipes* to materials that can be currently sourced. This exchange showed both models converged to very similar choices and that these models can adapt to unusual circumstances.

The ability to perform Choice seems to be hindered by alignment concerns such as the examples in the medical domain suggest. We also observe the very low precision of the responses, where even if the model has a solution that is a great fit to the problem, it still adds alternative solutions. This might be due to the goal of providing guidance to humans or simply to increase a minimum number of tokens to guarantee profitability. Further studies are needed to determine the ability of 2025's models to help society imagine novel solutions to its problems.

The main challenge faced to evaluate LLMs is they are designed for lengthy generations with low precision allowing for high recall of contents. The broad data used for training enables them to return correct answers to most questions, even if among wrong ones. Using any existing AI algorithm as baseline is challenging as AI algorithms typically execute reasoning tasks such as design, classification, prescription, configuration, and not generation for guidance to humans such as LLMs are designed and calibrated for alignment. Humans are likely poor baselines because when humans know something, they will provide answers with high precision, making their responses not suitable comparisons.

3. CBR-LLM Synergies

As discussed in the literature [4], there are many opportunities to explore synergies between CBR and LLMs. Based on this work, it seems that any hypothesis stating that 2025's models cannot *examine problems and creatively imagine and evaluate solutions to those problems* could be easily debunked. We did not demonstrate they can or that they engage in any reasoning, but we observed that 2025's models

---

[1] We asked about a hypothetical chest infection looking for a diagnosis in web-based Claude 3.7 Sonnet with extended thinking enabled and GPT-o3. Complete transcriptions at the github link provided in Section 2.

can answer any question competently albeit with low precision. Consequently, if society faces a novel problem, current models can help brainstorm alternatives but they will likely not indicate one optimal solution. The problem we have now is how LLMs could help CBR to *examine problems and creatively imagine and evaluate solutions to those problems*, and how CBR could help LLMs improve their precision.

Given their ability of executing Intelligence, Design, and Choice, as per Simon's model [5], LLMs seem to be a rich source of data to build CBR systems. Based on the analysis in this paper, we observed that LLMs have the capacity to produce generations that include the connections between problems and solutions. The question now is whether this data source could provide any potential benefits if used as a new knowledge container. Could this help CBR systems *examine problems and creatively imagine and evaluate solutions to those problems*? Or should CBR systems be simply used as an external module to LLMs to guarantee high precision? Or should CBR be used in this capacity of promoting high precision for selecting data to fine-tune new generations of LLMs?

4. Related Work

Classical decision theories, such as Simon's intelligence-design-choice framework [5] and Huber's extended problem-solving model [6], describe decision-making as a structured process of perceiving information, generating alternatives, and selecting strategies based on constraints. These foundational models offer a lens for analyzing whether LLMs exhibit comparable internal coherence in solving both goal-oriented tasks and generative responses that aim at general guidance.

Huang and Chang [17] review the literature on reasoning in LLMs and warn that fluent or coherent responses often obscure the absence of genuine inference. They argue for distinguishing between coherence and grounded reasoning, motivating evaluations that test not just outputs, but the intermediate structure of responses. Plaat et al. [18] synthesize recent work on reasoning and planning in agentic LLMs. They note a proliferation of agent tasks framed as reasoning, yet find little evidence of internal structure linking problem analysis to response generation. Their synthesis underlines the need for metrics that evaluate whether models internally coordinate their outputs with problem elements.

Huang et al. [19] assess LLM behavior in multi-agent games using the GAMA(γ)-Bench framework. Although they report strategic variation between models, their analysis interprets variation in behavioral output as signs of adaptive intelligence, without identifying whether models exhibit consistent procedures to integrate rules, goals, and context. This raises questions about what strategic performance reflects: reasoning, overfitting, or stylistic mimicry.

Schaeffer et al. [20] challenge the notion that recent performance gains indicate emergent reasoning abilities. They argue that what appears to be cognitive sophistication is better explained by quirks of benchmark design and scaling laws, casting doubt on the idea that high few-shot performance equates to structured inference.

Chen et al. [21] empirically examine the fidelity of chain-of-thought reasoning and find that models often produce explanations that do not reflect the process used to derive the final answer. Their results suggest that explanations may serve a rhetorical function, dissociated from problem-solving behavior. For this reason, we do not explore the concept of explanatory contents in this paper, and limit our analysis to what LLMs can reveal about connection between problems and solutions.

Gubelmann [22] argues from a Wittgensteinian standpoint that language models can produce context-appropriate responses without understanding in the propositional sense. Although such models exhibit linguistic competence, their outputs are not governed by belief-like internal states or processes, reinforcing skepticism about claims of reasoning. Min et al. [12] introduce FActScore to evaluate factual consistency in long-form outputs. Although focused on truthfulness, the framework highlights the discrepancies between fluent text and verifiable content, illustrating the challenges of attributing structured reasoning to surface-level explanations.

Mugleston et al. [23] explore whether LLMs can be said to possess knowledge. They propose that models encode a form of compressed, generative knowledge via statistical patterns, but do not exhibit reflective access or structured deliberation akin to human cognition. They discuss three forms of reasoning. A priori reasoning, as exemplified in Kantian philosophy, derives conclusions from premises independently of experience, relying on logical or conceptual necessity [23]. In contrast, transformer-based language models generate outputs by predicting the next token based on statistical regularities in training data. These models produce reasoning-like behavior through large-scale pattern recognition, without encoding explicit inference procedures [24, 17]. A third category includes hybrid approaches that support more structured forms of reasoning by incorporating additional mechanisms alongside the core transformer architecture. These approaches include chain-of-thought prompting, which introduces intermediate steps [24], as well as methods involving planning components, task decomposition, memory systems, and symbolic logic integration [21, 25, 26]. Together, these configurations form a class of systems referred to here as Large Reasoning Models, which are developed to produce more coherent, interpretable, and context related responses.

5. Conclusions and Future Work

This paper examines the question of whether a set of current LLMs (*i.e.*, Gemini 2.0 Flash, Gemini 1.5 Pro, Claude 3.7 Sonnet, GPT-4o, GPT-4.1, GPT-o3) can execute the second step in Simon's decision-making model [5]. This question serves to determine the possibility of such models to examine problems and imagine solutions for it. The results show that 2025's LLMs are able to brainstorm solutions to problems and some can provide indications of knowing how problems and solutions connect. Preliminary examination of whether these models can competently perform the Choice step of the said decision-making model suggests that it may depend on the domain as alignment efforts may preclude models from taking a stand. It is our conclusion that 2025's model can brainstorm solutions to problems but their responses are low in precision because they include multiple solutions.

This paper discusses whether this topic reveals novel opportunities to explore synergies between CBR and LLMs. Aware of the limitations of CBR systems to creatively explore novel solutions to previously unseen problems, we ask whether data from LLMs might become a source to add novel capabilities to CBR systems. The goal for any new developments would be to provide solutions with high precision, meeting this need left unmet by LLMs. Another direction is to have CBR modules help LLMs increase their precision. Ultimately, we would like to have intelligent systems that can help us solve previously unseen problems, those with which humans struggle the most.

Acknowledgments

The authors thank the reviewers for their excellent suggestions that improved this work. The first and second authors were funded by the Defense Advanced Research Projects Agency (DARPA), contract number FA8650-23-C-7317. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools to generate ideas for this work.

References

[1] M. Scheutz, V. Sarathy, From thinking to inventing, NSF 2026 Idea Machine Grand Prize Winner, 2019. URL: https://www.nsf.gov/about/history/big-ideas.
[2] R. C. Schank, Dynamic memory revisited, Cambridge University Press, 1999.

[3] M. M. Richter, R. O. Weber, Case-based reasoning: a textbook, Springer-Verlag: Berlin Heidelberg., 2013.

[4] K. Bach, R. Bergmann, F. Brand, M. Caro-Martínez, V. Eisenstadt, M. W. Floyd, L. Jayawardena, D. Leake, M. Lenz, L. Malburg, et al., Case-based reasoning meets large language models: A research manifesto for open challenges and research directions (2025).

[5] H. Simon, Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations, 4th ed., The Free Press, 1997.

[6] G. P. Huber, Managerial Decision Making, Scott, Foresman and Co., Glenview, IL, 1980.

[7] A. Asemi, A. Safari, A. A. Zavareh, The role of management information system (mis) and decision support system (dss) for manager's decision making process, International Journal of business and management 6 (2011) 164–173.

[8] J.-t. Huang, W. Wang, E. J. Li, M. H. Lam, S. Ren, Y. Yuan, W. Jiao, Z. Tu, M. R. Lyu, Who is chatgpt? benchmarking llms' psychological portrayal using psychobench, CoRR (2023).

[9] N. Milano, M. Ponticorvo, D. Marocco, Comparing human expertise and large language models embeddings in content validity assessment of personality tests, arXiv preprint arXiv:2503.12080 (2025).

[10] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1049–1065.

[11] S. Kambhampati, Can large language models reason and plan?, Annals of the New York Academy of Sciences 1534 (2024) 15–18.

[12] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi, Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 12076–12100.

[13] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics Doklady 10 (1966) 707–710.

[14] OpenAI, Learning to reason with large language models, https://openai.com/index/learning-to-reason-with-llms/, 2024. Accessed: 2024-05-01.

[15] G. DeepMind, Large language models self-discover reasoning structures, https://deepmind.google/research/publications/64816/, 2024. Accessed: 2024-05-01.

[16] Anthropic, Tracing the thoughts of a large language model, https://www.anthropic.com/research/tracing-thoughts-language-model, 2024. Accessed: 2024-05-01.

[17] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, arXiv preprint arXiv:2212.10545 (2023).

[18] A. Plaat, et al., Reasoning, planning, and acting in large language model agents: A survey, arXiv preprint (2025).

[19] J.-t. Huang, E. J. Li, M. H. Lam, et al., How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments, in: International Conference on Learning Representations (ICLR), 2025.

[20] J. Schaeffer, et al., Are emergent abilities of large language models a mirage?, Transactions on Machine Learning Research (TMLR) (2023).

[21] Y. Chen, J. Benton, et al., Reasoning models don't always say what they think, Anthropic (2024).

[22] S. Gubelmann, A loosely wittgensteinian conception of the linguistic understanding of large language models like bert and gpt, Philosophy and Technology (2023).

[23] J. Mugleston, V. H. Truong, et al., Epistemology in the age of large language models, Knowledge 5 (2025) 3.

[24] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, Transactions on Machine Learning Research (2022).

[25] M. Pink, Q. Wu, V. A. Vo, J. S. Turek, J. Mu, A. Huth, M. Toneva, Position: Episodic memory is the missing piece for long-term llm agents, arXiv preprint arXiv:2502.06975 (2025).

[26] Z. Yang, A. Ishay, J. Lee, Coupling large language models with logic programming for robust and general reasoning from text, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5186–5219.