

MaScQA: Investigating Materials Science Knowledge of Large Language Models

Mohd Zaki¹, Jayadeva², Mausam^{3,4}, N. M. Anoop Krishnan^{1,3}

¹Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi
110016, India

²Department of Electrical Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi
110016, India

³Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, Hauz Khas, New Delhi
110016, India

⁴Department of Computer Science & Engineering, Indian Institute of Technology Delhi, Hauz Khas,
New Delhi 110016, India

SUPPLEMENTARY INFORMATION

Supplementary Discussion

In addition to the collection of questions from 14 subdomains of materials science, the proposed dataset MaScQA has questions with lengths ranging from 9 words to 145 words (See Figure S1). Here, we will discuss the performance of GPT-4-CoT from the materials science domain perspective. The topics in Fig. 5 are arranged in decreasing order of the total number of questions in each category. The maximum percentage of incorrect questions is in questions under the *mechanical behavior* topic, where GPT-4-CoT has the second-worst performance. Out of 54 incorrectly answered questions, 34 are numerical questions. The questions where mistakes happened were based on concepts of the materials' stress-strain curve, fracture mechanics, and creep behavior. The second category with the most incorrect questions is *electrical behavior*, where incorrectly answered questions were related to battery cells, redox reactions, or identifying the potentials between the electrodes. The number of numerical questions answered wrong is 3 – 5 times that of the other types of questions. The *thermodynamics* category has a maximum number of questions, and quite a high percentage of incorrectly answered questions (~43%). The incorrect questions require understanding concepts of formation energy, specific heat, heat transfer, and chemical equations, to name a few, and solving complex equations correctly. The category of *atomic structure* has 41% incorrectly answered, mostly related to questions on the analysis of X-ray diffraction studies to identify the crystal structure of the materials. This reflects that LLMs are unable to correlate theoretical concepts with experimental outcomes. The category *magnetic behavior* has fewer questions (15), of which only nine are correct. The performance of LLMs in answering these questions reflects their inability to retrieve related concepts like magnetic moment and saturation magnetizations and avoid numerical errors. The next category is *transport phenomenon*, where the incorrectly answered questions (~38%) required understanding diffusion phenomenon and concepts of thermodynamics and battery cell reactions. In *phase*

transition, the incorrectly answered questions (~37%) are related to solving for the composition of different phases after the transitions and conditions required for phase transition.

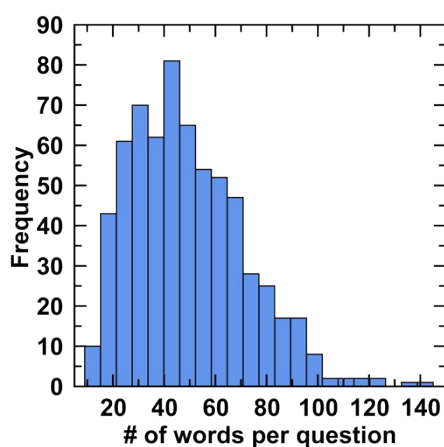


Figure S1. Visualising distribution of the number of words per question in MaScQA

Comparative analysis

Finally, to answer the third question raised in this work, i.e., what factors limit the performance of LLMs on MaScQA, we visualize the mistakes made by GPT-3.5-CoT and the solution provided by GPT-4-CoT. Fig. S2 shows one example where GPT-4-CoT yielded the correct answer. If we check the Wikipedia page for phase rule, the first expression is the one that is proposed as a solution by GPT-3.5. However, GPT-4 reaches the correct solution, which is also available on the same Wikipedia page. Although the dataset details on which these models are trained are unknown to the users, it is assumed that openly available sources like Wikipedia are a common dataset used by researchers while training such language models^{1,2}. Thus, it is interesting to note that while GPT-3.5 depicts a shallow understanding of concepts, GPT-4 can provide a deeper understanding based on the context.

At a constant atmospheric pressure, the number of phases, P which coexist in a chosen system at equilibrium, is related to the number of components, C in the system and the degree of freedom, F by
(A) $P+F=C-2$ (B) $P+F=C+2$ (C) $P+F=C+1$ (D) $P+F=C-1$

GPT-3.5-CoT: The Gibbs phase rule states that: $F = C - P + 2$ where F = degrees of freedom C = number of components P = number of phases Rearranging the equation: $P + F = C + 2$ Substituting the given values in this equation: $P + F = C + 2$ (at constant atmospheric pressure) Therefore, the answer is (B) $P+F=C+2$.

GPT-4-CoT: The relationship between the number of phases (P), the number of components (C), and the degree of freedom (F) in a system at equilibrium is described by the Gibbs Phase Rule. The equation for the Gibbs Phase Rule is: $P + F = C + 2 - \Phi$ where Φ is the number of non-compositional variables (such as pressure, temperature, etc.) that can be changed independently without altering the system's phase equilibrium. When the problem states that the atmospheric pressure is constant, it means that pressure is not a variable that can be changed independently since it remains the same throughout the system. Consequently, $\Phi = 1$ in this scenario. We are given that the atmospheric pressure is constant, so we can rewrite the equation as: $P + F = C + 2 - 1$ $P + F = C + 1$ Now we can see that the correct answer is (C) $P + F = C + 1$. Answer: [(C) $P + F = C + 1$]

Figure S2. Visualizing output of GPT models on a sample MCQ question.

The MATCH type questions require understanding different topics and then the ability to interlink them. An example of a matching question with the solution as per GPT-3.5-CoT and GPT-4-CoT is shown in Fig. S3. The scores in Table 2 indicated the exceptionally high performance of GPT-4 models in answering the matching-type question, which is more than two times the performance of GPT-3.5 models. It can be seen from the response of GPT-3.5-CoT that it is only able to determine the material properties required for the missile cone heads. Interestingly, GPT-3.5-CoT tries to arrive at the correct answer by eliminating the options. In contrast, GPT-4-CoT relied on understanding the topics and answering the questions after interrelating the previous information. This reinforces the idea that GPT-3.5 has a shallow understanding of the concepts.

Match the composite in Column I with the most suitable application in Column II.
 Column I : [(P) Glass fibre reinforced plastic, (Q) SiC particle reinforced Al alloy, (R) Carbon-carbon composite, (S) Metal fibre reinforced rubber]
 Column II: [(1) Missile cone heads, (2) Commercial automobile chasis, (3) Airplane wheel tyres, (4) Car piston rings, (5) High performance skate boards]
 Options: (A) P-4, Q-5, R-1, S-2 (B) P-3, Q-5, R-2, S-4 (C) P-5, Q-4, R-1, S-3 (D) P-4, Q-2, R-3, S-1

GPT-3.5-CoT: Composite | Application --- | --- P - Glass fibre reinforced plastic | 3 - Airplane wheel tyres Q - SiC particle reinforced Al alloy | 4 - Car piston rings R - Carbon-carbon composite | 1 - Missile cone heads S - Metal fibre reinforced rubber | 2 - Commercial automobile chasis
 Explanation: - Glass fiber reinforced plastic is known for its high strength-to-weight ratio and resistance to water and corrosion, making it suitable for airplane wheel tires (Option A and C eliminated).
 - SiC particle reinforced Al alloy is commonly used in high-performance pistons and piston rings because it has low thermal expansion, high thermal conductivity, and is resistant to wear and abrasion, making it suitable for car piston rings (Option C eliminated).
 - Carbon-carbon composites are known for their high strength and heat resistance, making them suitable for missile cone heads, which must withstand high temperatures and stresses during operation (Option B eliminated).
 - Metal fiber reinforced rubber is commonly used in commercial automobile chassis to improve the mechanical properties of rubber and is used to reduce noise and vibration (Option D eliminated). Therefore, the correct answer is (A) P-3, Q-5, R-1, S-2.

GPT-4-CoT: To correctly match the composite materials in Column I with their most suitable applications in Column II, we need to understand the properties of each composite material and the requirements for each application.

Column I: (P) Glass fibre reinforced plastic (GFRP) - Lightweight, strong, and resistant to corrosion.
 (Q) SiC particle reinforced Al alloy - High strength, stiffness, wear resistance, and thermal stability.
 (R) Carbon-carbon composite - High-temperature resistance, lightweight, and high strength.
 (S) Metal fibre reinforced rubber - Improved tensile strength, abrasion resistance, and flexibility.

Column II: (1) Missile cone heads - Require lightweight, heat-resistant, and high strength materials.
 (2) Commercial automobile chassis - Require high-strength, lightweight, and corrosion-resistant materials.
 (3) Airplane wheel tyres - Require materials with good tensile strength, abrasion resistance, and flexibility.
 (4) Car piston rings - Need high strength, wear resistance, and thermal stability.
 (5) High-performance skateboards - Require lightweight, strong, and corrosion-resistant materials.

Now let's try to match the materials with the applications: (P) GFRP - Its lightweight, strong, and corrosion-resistant properties make it suitable for high-performance skateboards. So, P matches with 5. (Q) SiC particle reinforced Al alloy - High strength, wear resistance, and thermal stability make this material suitable for car piston rings. So, Q matches with 4. (R) Carbon-carbon composite - High-temperature resistance, lightweight, and high strength properties make this material well-suited for missile cone heads. So, R matches with 1. (S) Metal fibre reinforced rubber - Improved tensile strength, abrasion resistance, and flexibility make this material suitable for airplane wheel tyres. So, S matches with 3. Thus, the correct matching is P-5, Q-4, R-1, S-3, which corresponds to option (C).

Figure S3. Visualizing output of GPT models on a sample MATCH question

An example of a MCQN question with is shown in Fig. S4. The GPT-3.5-CoT solution used the correct concept but made calculation errors, leading to an incorrect final answer. However, GPT-4-CoT used the correct concept and did not make calculation mistakes. It is observed in Table 2 that both GPT-4 and GPT-4-CoT achieve similar accuracy in answering MCQN questions. The red-coloured text in the GPT-3.5-CoT solution shows the source of the error, which led to an incorrect answer.

A 20 kN tensile load is applied axially to a steel bar of cross-sectional area 8 cm^2 and 1 m length. The Young's modulus of steel (E_{steel}) is 200 GPa, and of aluminium (E_{Al}) is 70 GPa. The Poisson's ratio (ν) can be taken as 0.3. When the same load is applied to an aluminium bar, it is found to give same elastic strain as the steel. Calculate the cross-sectional area of the aluminium bar.
 (A) 11.43 cm^2 (B) 14.93 cm^2 (C) 18.26 cm^2 (D) 22.86 cm^2

GPT-3.5-CoT: Given data, Force, $F = 20 \text{ kN} = 20 \times 10^3 \text{ N}$, Length, $L = 1 \text{ m}$, Cross-sectional area of steel bar, $A_{\text{steel}} = 8 \text{ cm}^2$, Young's modulus of steel, $E_{\text{steel}} = 200 \text{ GPa} = 200 \times 10^9 \text{ N/m}^2$, Young's modulus of aluminium, $E_{\text{Al}} = 70 \text{ GPa} = 70 \times 10^9 \text{ N/m}^2$, Poisson's ratio, $\nu = 0.3$
 Let ϵ_{steel} and ϵ_{Al} be the elastic strain in the steel and aluminium bars respectively. We also know that strain, $\Delta L/L = \epsilon_L/1 = \epsilon$ Where, ΔL is change in length and L is original length. Now, we can first calculate the stress in the steel bar as, $\sigma_{\text{steel}} = F/A_{\text{steel}}$ And then calculate the elastic strain in the steel bar as, $\epsilon_{\text{steel}} = \sigma_{\text{steel}}/E_{\text{steel}}$
 Using the fact that elastic strain is same for aluminium and steel, we can calculate the force that needs to be applied for the aluminium bar as follows, $\epsilon_{\text{Al}} = \frac{\sigma_{\text{Al}}}{E_{\text{Al}}}$, therefore, $\sigma_{\text{Al}} = \epsilon_{\text{Al}} \times E_{\text{Al}}$
 because $\epsilon_{\text{Al}} = \epsilon_{\text{steel}}$, therefore, $\sigma_{\text{Al}} = \epsilon_{\text{steel}} \times E_{\text{Al}}$ therefore $F_{\text{Al}} = \sigma_{\text{Al}} \times A_{\text{Al}}$
 because $F_{\text{Al}} = F = 20 \times 10^3 \text{ N}$ therefore $\sigma_{\text{Al}} \times A_{\text{Al}} = 20 \times 10^3 \text{ N}$
 Putting the value of σ_{Al} from above, we get $A_{\text{Al}} = \frac{20 \times 10^3 \text{ N}}{\epsilon_{\text{steel}} \times E_{\text{Al}}}$
 Now, we need to calculate the elastic strain in the steel bar. This can be done as follows, $\sigma_{\text{steel}} = F/A_{\text{steel}}$
 $\epsilon_{\text{steel}} = \sigma_{\text{steel}}/E_{\text{steel}}$, therefore $\epsilon_{\text{steel}} = F/A_{\text{steel}} \times E_{\text{steel}}$, therefore $\epsilon_{\text{steel}} = 20 \times 10^3 \text{ N} / 8 \times 10^{-4} \text{ m}^2 \times 200 \times 10^9 \text{ N/m}^2$
 therefore $\epsilon_{\text{steel}} = 1.25 \times 10^{-4}$
 Now, we can put the value of ϵ_{steel} in our equation for A_{Al} that we derived earlier to get the final answer.
 $A_{\text{Al}} = \frac{20 \times 10^3 \text{ N}}{\epsilon_{\text{steel}} \times E_{\text{Al}}} = \frac{20 \times 10^3 \text{ N}}{1.25 \times 10^{-4} \times 70 \times 10^9 \text{ N/m}^2}$
 $A_{\text{Al}} = 14.93 \text{ cm}^2$ (approx).
 Therefore, the cross-sectional area of the aluminium bar is approximately 14.93 cm^2 . Hence, the correct answer is option (B) 14.93 cm^2 .

GPT-4-CoT: To solve this problem, we first need to find the stress and strain in the steel bar when the load is applied. Then, we'll find the stress in the aluminium bar using the same strain, and finally, we'll calculate the cross-sectional area of the aluminium bar.
 Step 1: Find the stress in the steel bar Stress in the steel bar (σ_{steel}) is given by the formula: $\sigma_{\text{steel}} = \frac{\text{Load}}{\text{Area}}$
 Given load = 20 kN = 20,000 N, and the cross-sectional area of steel bar = $8 \text{ cm}^2 = 8 \times 10^{-4} \text{ m}^2$. So, $\sigma_{\text{steel}} = \frac{20,000}{8 \times 10^{-4}} = 2.5 \times 10^7 \text{ Pa}$
 Step 2: Find the strain in the steel bar Strain in the steel bar (ϵ_{steel}) is given by the formula: $\epsilon_{\text{steel}} = \frac{\sigma_{\text{steel}}}{E_{\text{steel}}}$ Given $E_{\text{steel}} = 200 \text{ GPa} = 200 \times 10^9 \text{ Pa}$, So, $\epsilon_{\text{steel}} = \frac{2.5 \times 10^7}{200 \times 10^9} = 1.25 \times 10^{-4}$
 Step 3: Find the stress in the aluminium bar Since the strain in the aluminium bar is the same as the strain in the steel bar: $\epsilon_{\text{Al}} = \epsilon_{\text{steel}}$
 Stress in the aluminium bar (σ_{Al}) is given by the formula: $\sigma_{\text{Al}} = E_{\text{Al}} \times \epsilon_{\text{Al}}$
 Given $E_{\text{Al}} = 70 \text{ GPa} = 70 \times 10^9 \text{ Pa}$, and $\epsilon_{\text{Al}} = 1.25 \times 10^{-4}$, So, $\sigma_{\text{Al}} = 70 \times 10^9 \times 1.25 \times 10^{-4} = 8.75 \times 10^6 \text{ Pa}$
 Step 4: Calculate the cross-sectional area of the aluminium bar Let the cross-sectional area of the aluminium bar be A . Stress in the aluminium bar is given by the formula: $\sigma_{\text{Al}} = \frac{\text{Load}}{\text{Area}}$
 From the above equation, we get: $\text{Area} (A) = \frac{\text{Load}}{\sigma_{\text{Al}}}$ Given load = 20,000 N, and $\sigma_{\text{Al}} = 8.75 \times 10^6 \text{ Pa}$, So, $A = \frac{20,000}{8.75 \times 10^6} = 2.2857 \times 10^{-3} \text{ m}^2$. Converting to cm^2 , we get: $A = 2.2857 \times 10^{-3} \times 10^4 = 22.857 \text{ cm}^2$. Therefore, the cross-sectional area of the aluminium bar is approximately 22.86 cm^2 . The correct option is (D) 22.86 cm^2 .

Figure S4. Visualizing the output of GPT models on an MCQN type of question

Now, we show the comparison of the solution by GPT-3.5-CoT and GPT-4-CoT on a sample NUM question in Fig. S5 related to platinum's crystal structure. Both models applied the correct concept. However, GPT-3.5-CoT made a calculation mistake in obtaining the interplanar distance “ d ”, which is highlighted in boldface and red colour in Fig S5. Calculation mistakes

are a known issue with such kinds of LLMs from the literature^{1,3-6} where similar order of accuracy was achieved on numerical questions solving tasks. The low accuracy of LLMs may also imply a lack of material science concepts previously observed in MCQ and MATCH-type questions in addition to calculation in capability.

The expected diffraction angle (in degrees) for the first order reflection from the (113) set of planes for face centered cubic Pt (lattice parameter = 0.392 nm) using monochromatic radiation of wavelength 0.1542 nm is

GPT-3.5-CoT: To solve this problem, we will use Bragg's law: $n\lambda = 2d\sin\theta$ where n is the order of diffraction (in this case, $n = 1$), λ is the wavelength of the radiation (0.1542 nm), d is the interplanar spacing, and θ is the diffraction angle we are trying to find. To find d for the (113) planes of face-centered cubic Pt, we can use the formula: $d = a / \sqrt{h^2 + k^2 + l^2}$ where a is the lattice parameter (0.392 nm), and h , k , and l are the Miller indices for the set of planes we are interested in (113): $d = 0.392 \text{ nm} / \sqrt{1^2 + 1^2 + 3^2} = 0.136 \text{ nm}$ Now we can use Bragg's law to find θ : $\sin\theta = n\lambda / 2d = (1)(0.1542 \text{ nm}) / (2)(0.136 \text{ nm}) = 0.566$ $\theta = \sin^{-1}(0.566) = 34.6^\circ$ Therefore, the expected diffraction angle for the first order reflection from the (113) set of planes for face centered cubic Pt using monochromatic radiation of wavelength 0.1542 nm is 34.6° . Answer: [34.6°]

GPT-4-CoT: To solve this problem, we'll use Bragg's Law, which relates the angle of reflection to the interplanar distance (d-spacing) and the wavelength of the radiation. Bragg's Law is given by: $n\lambda = 2d \sin\theta$ where: - n is the order of reflection (in this case, $n = 1$ for first order) - λ is the wavelength of the radiation (in this case, $\lambda = 0.1542 \text{ nm}$) - d is the interplanar distance - θ is the angle of reflection (the value we want to calculate) First, we need to find the interplanar distance (d-spacing) for the (113) set of planes in a face-centered cubic (FCC) lattice. To do this, we'll use the formula for d-spacing in a cubic lattice: $d = (a) / \sqrt{h^2 + k^2 + l^2}$ where: - a is the lattice parameter (in this case, $a = 0.392 \text{ nm}$) - h , k , and l are the Miller indices for the set of planes (in this case, $h = 1$, $k = 1$, and $l = 3$) Plug in the values into the d-spacing formula: $d = (0.392 \text{ nm}) / \sqrt{(1)^2 + (1)^2 + (3)^2}$ $d = (0.392 \text{ nm}) / \sqrt{1 + 1 + 9}$ $d = (0.392 \text{ nm}) / \sqrt{11}$ $d \approx 0.1181 \text{ nm}$ Now that we have the d-spacing, we can use Bragg's Law to find θ : $1 * (0.1542 \text{ nm}) = 2 * (0.1181 \text{ nm}) * \sin\theta$ Rearrange the equation to solve for $\sin\theta$: $\sin\theta = (0.1542 \text{ nm}) / (2 * 0.1181 \text{ nm})$ $\sin\theta \approx 0.6531$ Now, to find the angle θ , we need to take the inverse sine (arcsin) of the $\sin\theta$ value: $\theta = \arcsin(0.6531)$ $\theta \approx 40.687^\circ$ So, the expected diffraction angle for the first order reflection from the (113) set of planes in a face centered cubic Pt lattice using monochromatic radiation of wavelength 0.1542 nm is approximately 40.687° . Answer: [40.687]

Figure S5: Visualizing output of GPT models on a NUM-type question

References:

1. Hoffmann, J. *et al.* Training Compute-Optimal Large Language Models. Preprint at <https://doi.org/10.48550/arXiv.2203.15556> (2022).
2. Rae, J. W. *et al.* Scaling language models: Methods, analysis & insights from training gopher. *ArXiv Prepr. ArXiv2112.11446* (2021).
3. OpenAI. GPT-4 Technical Report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2023).

4. Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. Preprint at <https://doi.org/10.48550/arXiv.2302.13971> (2023).
5. Peng, B., Li, C., He, P., Galley, M. & Gao, J. Instruction Tuning with GPT-4. Preprint at <http://arxiv.org/abs/2304.03277> (2023).
6. Touvron, H. *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. Preprint at <https://doi.org/10.48550/arXiv.2307.09288> (2023).