

# MATRIX: STRESS-TESTING LLM REASONING IN MATERIALS SCIENCE

**Junkai Zhang\***

University of California, Los Angeles

**Xiaoxuan Wang**

University of California, Los Angeles

**Changquan Gu**

University of California, Los Angeles

**Yanqiao Zhu**

University of California, Los Angeles

**Dawei Zhou**

Virginia Tech

**Wei Wang**

University of California, Los Angeles

**Jingru Gan\***

University of California, Los Angeles

**Zian Jia**

Princeton University

**Jianpeng Chen**

Virginia Tech

**Mingyu Derek Ma**


University of California, Los Angeles


**Ling Li**

University of Pennsylvania

## ABSTRACT

Large Language Models (LLMs) have demonstrated remarkable abilities in scientific reasoning, yet their reasoning capabilities in materials science remain underexplored. To fill this gap, we introduce **MATRIX**, a comprehensive **M**aterials science **R**easoning and **I**ntelligence **e**Xamination comprising 1,340 problems that span the essential subdisciplines of materials science. **MATRIX** features a structured and fine-grained taxonomy that categorizes materials science questions into 6 primary fields and 31 sub-fields, and includes a three-tier difficulty classification based on the reasoning length required to solve each question. **MATRIX** provides detailed reference solutions enabling precise error analysis and incorporates multimodal reasoning through visual contexts in 315 questions. Evaluations of leading models reveal that even the highest-performing model, *Gemini-2.5-Pro*, achieves under 80% accuracy on college-level materials science questions, highlighting the complexity of **MATRIX**. Our systematic analysis of different reasoning strategies—basic chain-of-thought, tool augmentation, and self-correction—demonstrates that no single method consistently excels across all scenarios. We further analyze performance by difficulty level, examine trade-offs between efficiency and accuracy, highlight the challenges inherent in multimodal reasoning tasks, analyze failure modes across LLMs and reasoning methods, and evaluate the influence of retrieval-augmented generation. **MATRIX** thus establishes a comprehensive and solid benchmark for assessing and driving improvements in the scientific reasoning capabilities of LLMs within the materials science domain.

 **Dataset:** [huggingface.co/datasets/JunkaiZ/MATRIX](https://huggingface.co/datasets/JunkaiZ/MATRIX)

 **Code:** [github.com/Jun-Kai-Zhang/MATRIX.git](https://github.com/Jun-Kai-Zhang/MATRIX.git)

## 1 INTRODUCTION

Recent years have witnessed remarkable advancements of LLM reasoning abilities. From Chain of thought (Wei et al., 2022) to self-correction (Shinn et al., 2023) and tool-augmentation (Gou et al., 2023), the boundaries of LLM reasoning have expanded dramatically. What began with grade-school arithmetic calculations (Cobbe et al., 2021) has evolved to solving problems at the level of International Mathematical Olympiad (IMO) silver medalists (DeepMind, 2024). The o-series

---

\*Equal contribution.

model of OpenAI’s can even solve a substantial portion of frontier mathematical problems that would typically require hours of concentrated effort from expert mathematicians (OpenAI, 2025; Glazer et al., 2024).

Beyond LLMs’ notable achievements in mathematics, general scientific reasoning has emerged as a new area of interest, where solving problems requires a proper combination of reasoning and domain-specific knowledge (Truhn et al., 2023; Ma et al., 2024a;b). Scientific reasoning benchmarks reveal that LLMs suffer from identifying correct scientific assumptions and often demonstrate flawed understanding of scientific formulas and principles (Wang et al., 2023). Those findings indicate that scientific reasoning presents unique challenges to LLMs compared to pure mathematical questions. Therefore, numerous benchmarks have been proposed towards assessing LLM’s scientific reasoning capability, spanning from grade-school (Lu et al., 2022) to PhD-level (Feng et al., 2025) problems across domains (Huang et al., 2024a; Acharya et al., 2023).

Despite the abundance of scientific problem-solving benchmarks, LLMs’ reasoning abilities in materials science remain underexplored. Materials science occupies a unique position at the intersection of physics and chemistry, bridging fundamental science and engineering applications. This interdisciplinary field inherently relies on knowledge integration across multiple domains and requires complex reasoning capabilities. Existing reasoning benchmarks in materials science are limited by the lack of comprehensive evaluation and correct solutions (Zaki et al., 2024), or by the dependence on synthetic data generated by LLMs themselves, which introduces unavoidable noises (Alampara et al., 2024). In addition, none of the existing benchmarks adequately assesses the multimodal reasoning ability of LLM in material science.

To comprehensively evaluate LLMs’ reasoning abilities in materials science, we propose **MATRIX**, a benchmark comprising 1340 meticulously curated questions from 10 college-level textbooks spanning essential subdisciplines of materials science. All questions are open-ended to prevent model guessing while enabling objective assessment through rule-based judgment. For structured evaluation, **MATRIX** constructs a comprehensive and fine-grained taxonomy with 6 primary fields (Materials, Properties, Structures, Fundamental Mechanisms, Processes, Failure Mechanisms) and 31 sub-fields that capture materials science’s interdisciplinary nature, enabling assessment of reasoning abilities on specific domains. In addition, questions are classified into three difficulty levels based on reasoning length required to solve the question, with 50.7% easy, 29.1% medium, and 20.1% hard questions. The 270 hard questions require long solving process, deliberately challenging models’ complex reasoning capabilities. Detailed solutions to 944 of the questions are included to facilitate error categorization and process-level evaluations. The benchmark also incorporates 315 questions with visual contexts to assess multimodal reasoning abilities.

The o-series models from OpenAI, such as o4-mini, along with Gemini-2.5-Pro, DeepSeek-R1, GPT-5, Claude-4-Sonnet, and Qwen3-235b-a22b-thinking, represent a new class of LLMs that exhibit complex reasoning by generating extended intermediate outputs before producing final answers. These models are informally referred to as *thinking models* or reasoning models, distinguishing them from traditional LLMs like GPT-4.1, Claude-3.7, DeepSeek-V3, Llama-4-Maverick, and Gemini-2.0-Flash, classified as *non-thinking models* (Chen et al., 2025). We conduct extensive experiments on **MATRIX** to evaluate and compare the reasoning capabilities of these six thinking models against five non-thinking models in materials science problem solving. In addition, we also evaluate the effectiveness of self-correction and tool-augmentation (i.e., integration of Python code) on non-thinking models in addition to the basic CoT. Our results indicate that while Gemini-2.5-Pro lead with approximately 77% accuracy, the best-performing non-thinking model, Llama-4-Maverick, achieves a comparable 71%. However, none of the techniques—basic CoT, self-correction, or tool-augmentation—consistently outperforms the others across all models, demonstrating that effectiveness depends significantly on the base model.

Our systematic analysis of LLM reasoning capabilities examine multiple dimensions: difficulty levels, reasoning efficiency, multimodal reasoning, and failure patterns. The key findings from our analysis include: (1) thinking models’ performance is insensitive to question difficulty, suggesting that they better handle reasoning-intensive tasks; (2) performance improves with longer outputs, establishing a clear efficiency-accuracy trade-off frontier; (3) image-included questions lead to poorer performance in multimodal models compared to text-only questions on the same LLMs, highlighting the inherent challenges of multimodal reasoning; (4) by categorizing incorrectly answered responses into predefined error types, we discovered that all tested models suffer from errors based on domain

knowledge inaccuracies and question comprehension failures. Although the three reasoning methods are capable of reducing specific types of errors, they may concurrently amplify other types of errors; (5) our case study suggests that web search-based RAG may have limited effectiveness in reducing knowledge-based errors and could potentially contribute to increased hallucination rates.

Our contributions are listed as follows:

- We introduce **MATRIX**, a comprehensive and challenging materials science reasoning benchmark comprising 1340 expert-curated questions from college-level textbooks across essential subdisciplines, featuring a structured taxonomy of 6 primary fields and 31 sub-fields, three-tier difficulty classification, detailed solutions for 944 questions, and 315 questions with visual contexts for multimodal reasoning evaluation.
- We benchmark SOTA LLMs, including six thinking models and five non-thinking models. Additionally, we enhance the non-thinking models with three popular reasoning methods. This provides the most comprehensive evaluation and comparison of reasoning capabilities in materials science across different models and methods.
- We present a comprehensive multi-dimensional analysis of LLM reasoning capability across difficulty levels, reasoning efficiency, accuracy trade-offs, multimodal reasoning capabilities, and failure patterns. We also assess the capability of small language models on **MATRIX**. We additionally conduct a case study exploring the influence of RAG on scientific reasoning in materials science. This thorough evaluation establishes a foundation for future improvements in scientific reasoning models.

## 2 RELATED WORK

As LLMs continue to develop reasoning abilities, solving scientific problems is considered a fundamental dimension and has been the focus of numerous benchmarks. GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), along with a series of benchmarks (Mirzadeh et al., 2024) evaluated the mathematical abilities of language models. With the emergence of multimodal LLMs, MathVista (Lu et al., 2023) further includes visual contexts to benchmark the multimodal reasoning abilities. With the growth of reasoning capabilities, competitive level questions like OlympiadBench (He et al., 2024) and PutnamBench (Tsoukalas et al., 2024), and advanced graduate-level math like Frontier Math (Glazer et al., 2024) and HARDMATH (Fan et al., 2024) set new standards for reasoning models.

Beyond mathematics, natural science questions involve not only reasoning but also domain knowledge, thus incentivizing increased interest, particularly in chemistry, physics, and biology (Welbl et al., 2017; Lu et al., 2022; Rein et al., 2024). SciBench (Wang et al., 2023), MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b) covers college-level scientific question solving requires both domain knowledge and sophisticated reasoning. OlympicArena (Huang et al., 2024b) contributes Olympiad-level, multimodal problems across seven scientific fields, and SuperGPQA (Du et al., 2025) further expands coverage to 285 graduate-level disciplines. Besides problem solving, SciEval (Sun et al., 2024) and SciKnowEval (Feng et al., 2024) evaluate multi-level capabilities of LLM in scientific domain. In addition to those general natural scientific reasoning benchmarks, a series of works Acharya et al. (2023); Li et al. (2025a) focus on specific domains. PhysReason (Zhang et al., 2025), PHYSICS (Feng et al., 2025), and MM-PhyQA (Anand et al., 2024) specialize on the physical questions; ChemEval (Huang et al., 2024a) benchmarks chemistry abilities; Sarwal et al. (2024) for Bioinformatics; Meshram et al. (2024) for electronics; and MaScQA (Zaki et al., 2024) for materials science question answering.

## 3 DATASET

### 3.1 DATA COLLECTION AND PROCESSING

For our materials science benchmark dataset, we curated a collection of problems from textbooks across multiple sub-fields. We selected widely-adopted undergraduate and graduate textbooks that include both comprehensive references (like “The Science and Engineering of Materials”) and specialized resources focusing on specific domains (such as “Electronic, Magnetic, and Optical

Materials”). The choice of textbooks was guided and validated by materials science experts. We first identified the major subfields of materials science and then selected textbooks in these areas that provide exercise solutions and are accessible online. These sources collectively provide diverse problem types that cover the breadth of materials science concepts. The full list of textbooks are presented in the Table 1.

Table 1: Source Textbooks Used for Question-answer Collection

Textbook	# QAs
Intro. to Materials Science for Engineers (Shackelford, 2016)	349
The Science & Eng. of Materials (Askeland et al., 2003)	287
Materials Science & Eng.: An Intro. (Callister Jr and Rethwisch, 2020)	61
Fund. of Materials Science & Eng. (Callister and Rethwisch, 2016)	393
Mechanical Behavior of Materials (Hosford, 2010a)	83
Electronic, Magnetic, & Optical Materials (Fulay and Lee, 2016)	72
Materials & Process Selection for Eng. Design (Frag, 2020)	27
Fund. of Ceramics (Barsoum, 2019)	29
Physical Metallurgy (Hosford, 2010b)	27
Polymer Science & Technology (Fried, 2014)	12
<b>Total</b>	<b>1340</b>

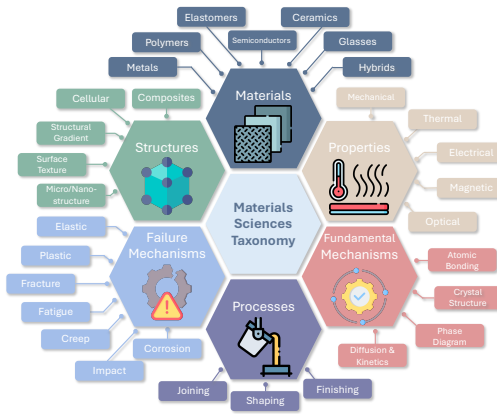


Figure 1: Taxonomy of **MATRIX** Materials Science QAs.

We used Mistral optical character recognition (OCR) (Mistral AI Team, 2025) to digitize both textual and visual content of these textbooks. Then we implemented a parsing algorithm to identify the example problems and solutions from the digital copies. Each question-answer pair was structured into a standardized format. Following the initial extraction, each entry was manually reviewed and corrected to ensure accuracy and completeness. We applied strict filtering criteria, retaining only questions with determinate answers in the form of numerical values or formulas.

### 3.2 DATASET STATISTICS

Our benchmark comprises 1340 question-answer pairs structured in a standardized format. Each entry contains fields for question text, solution text, associated images, difficulty level, domain classification, and problem type. Questions are categorized as either numerical or formula type according to the answer, with 92.4% being numerical and 7.6% requiring formula derivation, 315 questions (23.5%) include images.

### 3.3 TAXONOMY CLASSIFICATION

We developed a comprehensive hierarchical taxonomy to systematically categorize questions across fundamental materials science domains. Our taxonomy design was informed by established materials science curricula and reference texts, including Shackelford (2015); Shackelford et al. (2016); Ashby et al. (2019). The taxonomy framework reflects both the traditional organization of materials science education and the critical concepts that underpin fundamental mechanisms at different length scales. The taxonomy consists of six primary fields, each containing detailed subcategories:

1. **Materials:** Metals, Polymers, Elastomers, Semiconductors, Ceramics, Glasses, Hybrids
2. **Properties:** Mechanical, Thermal, Electrical, Magnetic, Optical
3. **Structures:** Composites, Cellular, Structural Gradient, Surface Texture, Micro/Nano-structure
4. **Fundamental Mechanisms:** Atomic Bonding, Crystal Structure, Phase Diagram, Diffusion & Kinetics

5. **Processes:** Joining, Shaping, Finishing
6. **Failure Mechanisms:** Elastic, Plastic, Fracture, Fatigue, Creep, Impact, Corrosion

Figure 1 shows our manually developed taxonomy that covers the domains of QAs collected in **MATRIX**. The taxonomy consists of six primary fields: Materials, Properties, Structures, Fundamental Mechanisms, Processes, and Failure Mechanisms, each containing detailed subcategories. This multi-dimensional classification scheme enables us to capture the interdisciplinary nature of materials science problems, where a single question might span multiple domains. Our taxonomy not only provides a nuanced understanding of the dataset composition but also enables targeted evaluation of model performance across specific subfields and their intersections, offering insights into how AI models handle different aspects of materials science knowledge. Additional details of dataset can be found in Appendix A.

### 3.4 DIFFICULTY CLASSIFICATION

We implemented a three-tier classification from easy, medium to hard, to assess question difficulty. Difficulty is assessed with response lengths from `Claude-3.7-Sonnet`, which classifies questions into Easy (50.7%), Medium (29.1%), and Hard (20.1%) categories based on the length of model responses required to solve them. This distribution provides a balanced representation across difficulty levels, while differentiating questions in terms of the knowledge and reasoning length required to derive a correct solution. We validated this length-based approach across multiple models and consistently observed that accuracy decreases while response length increases with difficulty, confirming the reliability of this assessment method. To verify the robustness of our classification, we use step-count analysis based on the judgment of `Gemini-2.0-Flash` for solution steps required to solve each question, along with additional pattern-based and KNN-based validation methods. Details of the validation approaches are discussed in Appendix A.3.

### 3.5 DATA LEAKAGE DETECTION

Data leakage is a key concern when constructing reliable benchmarks. We use the method proposed in Xu et al. (2024) to detect potential data leakage in our benchmark. Because this method is restricted to locally served models and the released implementation is not optimized for very large models, we evaluate a set of small language models: `Qwen-2.5-7B`, `Qwen-2.5-32B`, `Gemma-3-4B-it`, `Gemma-3-12B`, and `Gemma-3-27B`. The results, reported in Table 2, show N-gram accuracy and PPL only slightly above 1 for all models, indicating minimal overlap with training data and suggesting that **MATRIX** is free of data leakage.

Table 2: Data Leakage Detection on Small Language Models.

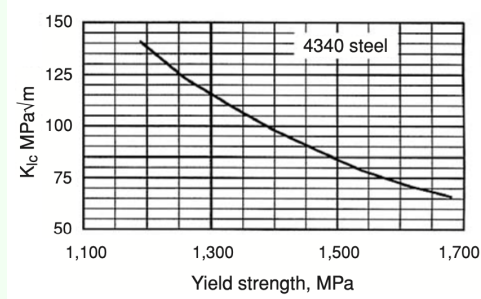
Model	N-gram Acc.	PPL
<code>Qwen-2.5-Instruct-7B</code>	1.28	1.07
<code>Qwen-2.5-Instruct-32B</code>	1.22	1.06
<code>Gemma-3-4B-it</code>	1.04	1.08
<code>Gemma-3-12B-it</code>	1.06	1.10
<code>Gemma-3-27B-it</code>	1.09	1.07

#### An Example Question from **MATRIX**

##### Question

A support is to be made from 4340 steel. The steel may be tempered to different yield strengths. The correlation between strength and toughness is shown in the figure. Assume that  $f = 1.1$  for the support geometry and that nondestructive inspection can detect all edge cracks of size  $a = 2$  mm or larger. If a 4340 steel of  $K_{Ic} = 120 \text{ MPa}\sqrt{\text{m}}$  is used, what is the largest stress that will not cause either yielding or fracture?

##### Image

**Solution**

From the figure, for  $K_{Ic} = 120 \text{ MPa}\sqrt{\text{m}}$  and  $Y = 1275 \text{ MPa}$ , using linear elastic fracture mechanics equation

$$\sigma_f = K_{Ic} / \left[ f(\pi a)^{1/2} \right],$$

$$\sigma_f = \frac{K_{Ic}}{f\sqrt{\pi a}} = \frac{120 \text{ MPa}}{1.1\sqrt{\pi(0.002 \text{ m})}} = 1376 \text{ MPa}.$$

This is higher than the yield strength, so it will fail by yielding when  $\sigma = 1275 \text{ MPa}$ .

**Answer** 1275

**Unit** MPa

**Primary Category** Materials: Metals

**Categories**

Materials: Metals

Properties: Mechanical

**Type** NUM

**Difficulty Level** easy

## 4 EXPERIMENTS

### 4.1 MODELS AND METHODS

For proprietary models, we evaluate GPT-4.1 (OpenAI, 2025), Claude-3.7-Sonnet (Anthropic, 2025), Gemini-2.0-Flash (Google DeepMind, 2024), and the thinking models o4-mini (OpenAI, 2024), Gemini-2.5-Pro (Google DeepMind, 2025), GPT-5 (OpenAI, 2025), Claude-Sonnet-4 (Anthropic, 2025); for open-weight models, we evaluate DeepSeek-V3 (Liu et al., 2024), llama-4-maverick (Meta AI, 2025), and the thinking models DeepSeek-R1 (Guo et al., 2025), Qwen3-235b-a22b-thinking (Yang et al., 2025). Among these models, GPT-4.1, Claude-3.7-Sonnet, Gemini-2.0-Flash, o4-mini, Gemini-2.5-Pro, GPT-5, and Claude-Sonnet-4 support visual inputs.

For non-thinking models, we adopt three prompting methods: basic CoT, self-correction, and tool-augmentation. The self-correction methods follows Huang et al. (2023); Kim et al. (2023); Shinn et al. (2023), invoking 3 rounds of conversation with the model: (1) the initial response, (2) detecting issues in the initial attempt, and (3) revising the initial attempt based on the detected problem. The tool-augmentation method prompts the model to generate Python code, executes it using a code interpreter (Gou et al., 2023; Yang et al., 2024), and derives the final answer based on the execution results. The detailed prompts are provided in the Appendix B.1.

### 4.2 EVALUATION

The correctness of the output answers is evaluated using a hybrid approach that combines rule-based evaluation and LLM-based evaluation. We adapt the rule-based evaluation system from Qwen-2.5 Math (Yang et al., 2024). Following the previous works (Methani et al., 2020; Gupta et al., 2024), we apply a relaxed numerical tolerance of 5% to account for approximation errors in calculations and image recognition. To address the limitations of rule-based systems in handling complex formulas

Table 3: Experimental Results in Terms of Accuracy Score (%) on **MATRIX**(questions w/o images). **Bold** indicates the best performance, and Underline indicates the second best.

Model	Fail. Mech.	Fund. Mech.	Materials	Processes	Properties	Structure	Overall
<i>Non-Thinking Models</i>							
<b>Claude-3.7-Sonnet</b>	65.66	65.97	65.89	63.64	64.84	68.74	67.32
+Correction	66.79	63.87	67.11	64.65	65.42	69.51	68.00
+Tool	<b>72.08</b>	66.18	<u>70.75</u>	64.65	<b>70.89</b>	<u>72.35</u>	<u>71.51</u>
<b>DeepSeek-V3</b>	62.64	61.97	65.67	63.64	63.26	66.89	66.15
+Correction	67.17	60.50	62.91	63.64	63.26	65.46	64.39
+Tool	61.51	59.03	62.69	62.63	57.93	64.15	62.44
<b>Gemini-2.0-Flash</b>	60.75	55.04	59.71	52.53	58.36	60.00	59.90
+Correction	59.62	59.24	61.92	51.52	59.22	63.28	62.34
+Tool	67.55	65.97	68.65	69.70	68.30	70.49	69.46
<b>GPT-4.1</b>	65.66	<b>68.91</b>	70.42	61.62	67.58	71.80	70.73
+Correction	66.04	65.13	68.10	57.58	65.56	69.29	68.00
+Tool	63.02	62.18	61.92	55.56	60.81	62.62	61.66
<b>Llama-4-Maverick</b>	69.06	<b>68.91</b>	<b>71.30</b>	<b>72.73</b>	<u>69.16</u>	<b>73.11</b>	<b>71.61</b>
+Correction	<u>69.43</u>	66.18	69.87	<u>70.71</u>	68.16	71.15	69.95
+Tool	68.30	63.24	68.21	67.68	65.27	68.85	68.20
<i>Thinking Models</i>							
<b>Claude-4-Sonnet</b>	58.49	52.52	54.86	56.57	52.31	54.64	54.44
<b>DeepSeek-R1</b>	71.70	71.43	<u>73.84</u>	<u>74.75</u>	72.62	<u>75.30</u>	73.95
<b>Gemini-2.5-Pro</b>	<b>78.49</b>	<b>76.89</b>	<b>77.15</b>	<b>75.76</b>	<b>74.50</b>	<b>78.69</b>	<b>77.37</b>
<b>Qwen3-235B</b>	<u>73.58</u>	69.96	71.96	68.69	70.17	73.33	72.10
<b>GPT-5</b>	67.17	64.71	65.34	67.68	62.97	65.57	64.88
<b>o4-mini</b>	72.08	<u>73.32</u>	73.73	69.70	<u>72.91</u>	74.97	<u>74.34</u>

Table 4: Experimental Results in Terms of Accuracy Score (%) on **MATRIX** (questions w/ images). **Bold** indicates the best performance, and Underline indicates the second best.

Model	Fail. Mech.	Fund. Mech.	Materials	Processes	Properties	Structure	Overall
<b>Claude-3.7-Sonnet</b>	29.76	37.13	37.31	40.62	34.45	34.51	34.60
<b>Claude-Sonnet-4</b>	30.95	<u>40.12</u>	39.18	<u>46.88</u>	35.89	<u>38.73</u>	37.46
<b>Gemini-2.0-Flash</b>	25.00	32.34	26.49	28.12	27.27	26.41	26.03
<b>Gemini-2.5-Pro</b>	<u>39.29</u>	<u>40.12</u>	<u>42.16</u>	31.25	<u>41.63</u>	<u>38.73</u>	<u>39.05</u>
<b>GPT-5</b>	<b>42.86</b>	<b>53.89</b>	<b>49.63</b>	<b>59.38</b>	<b>46.89</b>	<b>50.00</b>	<b>48.89</b>
<b>o4-mini</b>	33.33	40.72	37.69	43.75	36.36	37.32	37.14

and equations, we supplement this approach with Gemini-2.0-Flash for formula-type questions. The LLM’s judgment serves as the final determinant of correctness for these complex mathematical expressions. The performance in terms of accuracy score of all models on text-only questions is presented in Table 3, and the performance of multimodal models on images-included questions is presented in Table 4.

#### 4.3 RESULTS

**Observation 1.** Among non-thinking models, Llama-4-Maverick achieves the best overall accuracy (71.61%) under the basic chain-of-thought (CoT) setting. GPT-4.1 ranks second (70.73%) in the basic CoT category, although its performance decreases when tools are introduced. Claude-3.7-Sonnet shows relatively lower accuracy with basic CoT (67.32%), but improves to 71.51% with tool integration, becoming the second-best performer in the tool-augmented setting. Gemini-2.0-Flash has bad accuracy under the basic CoT condition (59.90%) but substantially improves with tool use, reaching 69.46%. For thinking models, Gemini-2.5-Pro attains the best results overall with 77.37%, surpassing all other models. DeepSeek-R1 is the strongest among open-weight

thinking models with 73.95%, closely followed by Qwen3-235B (72.10%). These results indicate that the performance gap between open-weight and proprietary models is narrowing.

**Observation 2.** No single prompting method demonstrates consistently superior performance across all models. The performance improvements achieved through tool-augmentation varies significantly between models: Claude-3.7 and Gemini-2.0-Flash show substantial increases in overall performance, GPT-4.1, DeepSeek-V3, and Llama-4-Maverick exhibits performance degradation. The self-correction technique generally decreases performance across most models, converting more correct answers to incorrect ones than vice versa. Only Gemini-2.0-Flash shows substantial performance improvements under this approach.

**Observation 3.** In the multimodal evaluation, GPT-5 delivers the strongest performance, achieving the highest overall accuracy (48.89%) and leading across all individual categories. Gemini-2.5-Pro ranks second overall with 39.05%. Claude models exhibit moderate performance, with Claude-Sonnet-4 (37.46%) slightly surpassing Claude-3.7-Sonnet (34.60%). The o4-mini model achieves a comparable score (37.14%) to Claude-Sonnet-4, but remains behind GPT-5 and Gemini-2.5-Pro. These results highlight the superiority of GPT-5 in handling multimodal reasoning tasks.

#### 4.4 PERFORMANCE OF SMALL LANGUAGE MODELS ON MATRIX

Small language models (SLMs) have recently become promising for agentic AI (Belcak et al., 2025). To assess their performance in the science domain, we evaluate the Qwen-2.5 family (Qwen et al., 2025) and the Gemma-3 family (Team et al., 2025) on **MATRIX**; the results are reported in Table 5. Overall, SLMs still struggle with knowledge-intensive scientific reasoning: even the best-performing model, Qwen-2.5-Instruct-32b, achieves only 33.27% accuracy. Performance is strongly correlated with model size, and the ranking of SLMs largely follows the ranking of their parameter scales.

Table 5: Experimental Results in Terms of Accuracy Score (%) on **MATRIX**(questions w/o images) for Small Language Models. **Bold** indicates the best performance, and Underline indicates the second best.

Model	Fail. Mech.	Fund. Mech.	Materials	Processes	Properties	Structure	Overall
Gemma-3-4b-it	4.15	6.30	6.73	4.04	5.76	6.99	7.12
Gemma-3-12b-it	12.45	7.98	13.13	5.05	12.68	13.55	13.76
Gemma-3-27b-it	<u>32.83</u>	<u>24.16</u>	<u>30.79</u>	<u>17.17</u>	<u>29.97</u>	<u>31.26</u>	<u>30.93</u>
Qwen2.5-7b-instruct	15.85	15.34	17.99	10.10	18.01	18.47	18.73
Qwen2.5-32b-instruct	<b>36.98</b>	<b>28.78</b>	<b>33.33</b>	<b>20.20</b>	<b>32.28</b>	<b>33.22</b>	<b>33.27</b>

## 5 ANALYSIS

### 5.1 PERFORMANCE ACROSS DIFFICULTY LEVELS

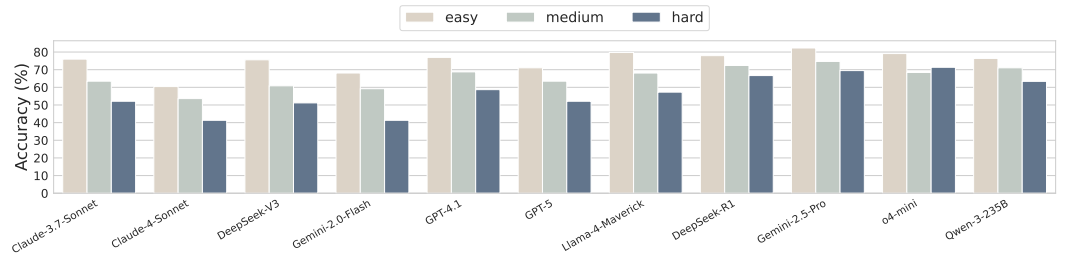


Figure 2: The Performance of LLMs across Difficulty Levels.

The accuracy scores of different models across difficulty levels are shown in Figure 2. Most models exhibit expected performance degradation patterns with increasing difficulty, suggesting that complex

reasoning process prevent them from reaching correct answers. o4-mini shows an interesting pattern: its accuracy on hard questions is not lower than on medium questions. This pattern may suggest that, for this relatively small-scale thinking model, the main difficulty might not lie in the length of reasoning, but rather in domain knowledge.

## 5.2 EFFICIENCY V.S. ACCURACY

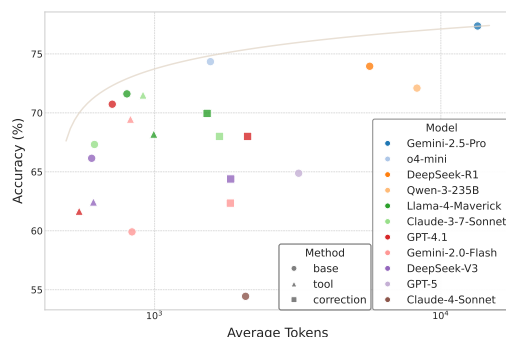


Figure 3: The Average Output Length v.s. The Accuracies.

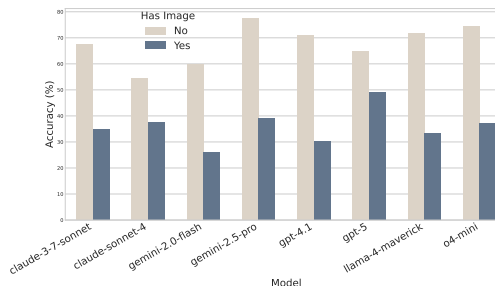


Figure 4: The Performance Comparison of MLLM between Questions w and w/o Images.

Thinking models often generate highly verbose outputs. This verbosity frequently involves branching, backtracking, error validation, and correction (Yeo et al., 2025), which, although beneficial for arriving at correct results, may compromise efficiency. This underscores a fundamental trade-off between reasoning accuracy and efficiency.

Figure 3 illustrates the relationship between performance and output length by showing token usage across different models and methods, with the boundary line representing the reasoning efficiency frontier. When using basic CoT prompting, thinking models consume significantly more tokens while achieving superior performance compared to non-thinking models. Self-correction prompting substantially increases output length without consistently improving performance—sometimes even degrading results. In contrast, tool augmentation provides a more economical approach, requiring minimal additional tokens while boosting performance across many models.

## 5.3 PERFORMANCE DROP DUE TO VISUAL CONTEXT

Image-included questions are significantly more challenging than text-only questions for multimodal LLMs, with a significantly lower accuracy scores, as presented in Figure 4. We identified two major sources of error in questions involving visual context: (i) many images in materials science are inherently three-dimensional—such as lattice cells or atomic arrangements—which challenge the spatial reasoning abilities of multimodal LLMs; (ii) many figures are diagrams or plots that require models to extract numerical values precisely, a task that remains difficult for current multimodal LLMs.

## 5.4 FAILURE PATTERN

To investigate the challenges LLMs face in solving materials science problems, we first manually reviewed a sample of incorrect responses to identify and define five major error categories: problem comprehension deficiencies, domain knowledge gaps, flawed solution strategies, calculation inaccuracies, and hallucinated content. We then employed Gemini-2.0-Flash to automatically categorize all incorrect responses into these five categories, covering all non-thinking models and prompting methods on text-only questions with reference solutions. When multiple errors exist, we classify them into the first appearing one in the solution. Detailed prompts and definitions for each category are provided in Appendix C.2.

The error rates across categories are presented in Figure 5. These findings reveal consistent patterns across all models, with deficiencies in domain knowledge and question comprehension representing

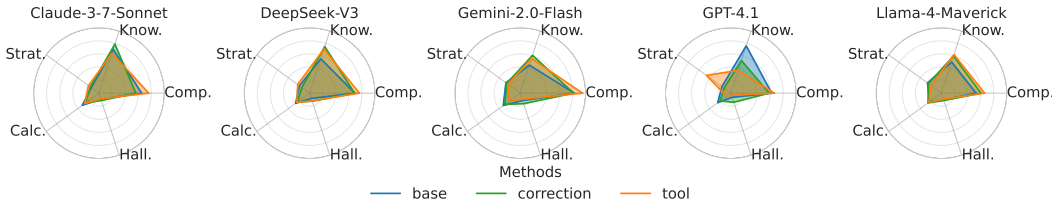


Figure 5: Error Categorization for Non-thinking models. Types of Errors Are Problem Comprehension Deficiencies (Comp.), Domain Knowledge Gaps (Know.), Flawed Solution Strategies (Strat.), Calculation Inaccuracies (Calc.), and Hallucinated Content (Hall.).

the most critical limitations—exceeding even calculation errors. While errors caused by hallucinations are still present, they occur less frequently than other error types. As expected, tool-augmentation methods reduced numerical errors across all models, with the most significant improvements observed in Gemini-2.0-Flash. Self-correction methods, on the other hand, did not provide consistent improvements across any of the tested models in any error category.

### 5.5 WEB SEARCH-BASED RETRIEVAL AUGMENTED GENERATION: A CASE STUDY

Retrieval Augmented Generation (RAG) has long been regarded as an effective approach to enhance model performance in scientific domains where specialized knowledge is necessary for completing tasks (Lála et al., 2023; Li et al., 2025b). To verify this approach on materials science reasoning tasks, we conducted a case study using DeepSeek-V3 on **MATRIX**. We implemented RAG through web searching: given a question, the LLM formulates a search query, retrieves up to five most relevant results from the Tavily API, summarizes the most useful information, and appends this to the original question. The failure pattern when using RAG is presented in Figure 6. Surprisingly, RAG does not reduce knowledge-related errors but instead improves problem comprehension. We hypothesize that web searching doesn’t consistently retrieve correct and useful information, thus fail to reliably enhance knowledge accuracy and occasionally even inducing hallucination. However, the additional contextual information may help the model better comprehend questions and identify relevant information for solving them. Specific examples for both cases can be found in Appendix C.3.



Figure 6: The Error Categories Counts of Basic CoT and RAG.

## 6 CONCLUSION AND LIMITATION

In this work, we present **MATRIX**, a benchmark comprising 1340 college-level materials science questions spanning all essential subdisciplines. We evaluate state-of-the-art thinking and non-thinking models on **MATRIX**, employing three different reasoning methods for non-thinking models. Our results reveal significant performance discrepancies among LLMs on materials science reasoning tasks and highlight the varying effectiveness of different reasoning approaches. We also evaluate small language models and find that they still struggle with knowledge-intensive scientific reasoning. We further analyze model performance across multiple dimensions: difficulty levels, reasoning efficiency, multimodal reasoning capabilities, failure patterns, and retrieval-augmented generation (RAG). This comprehensive analysis enhances our understanding of model performance and establishes a foundation for further improvements in materials science reasoning capabilities.

## REFERENCES

- Anurag Acharya, Sai Munikoti, Aaron Hellinger, Sara Smith, Sridevi Wagle, and Sameera Horawalavithana. Nuclearqa: A human-made benchmark for language models for the nuclear domain. *arXiv preprint arXiv:2310.10920*, 2023.
- Nawaf Alampara, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, Mara Schilling-Wilhelmi, NM Anoop Krishnan, and Kevin Maik Jablonka. Macbench: A multimodal chemistry and materials science benchmark. 2024.
- Avinash Anand, Janak Kapuriya, Apoorv Singh, Jay Saraf, Naman Lal, Astha Verma, Rushali Gupta, and Rajiv Shah. Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 53–64. Springer, 2024.
- Anthropic. Claude sonnet 4. <https://www.anthropic.com/claude/sonnet>, 2025. Accessed: 2025-09-24.
- Anthropic. Claude 3.7 sonnet and claude code. <https://www.anthropic.com/news/claude-3-7-sonnet>, February 2025. Accessed: 2025-05-15.
- Michael F. Ashby, Hugh Shercliff, and David Cebon. *Materials: Engineering, Science, Processing and Design*. Butterworth-Heinemann, Oxford, UK, 4 edition, 2019.
- Donald R Askeland, Pradeep Prabhakar Phulé, Wendelin J Wright, and DK Bhattacharya. *The science and engineering of materials*. 2003.
- Michel Barsoum. *Fundamentals of ceramics*. CRC press, 2019.
- Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*, 2025.
- William D. Callister and David G. Rethwisch. *Fundamentals of Materials Science and Engineering: An Integrated Approach*. Wiley Global Education, 5 edition, 2016. ISBN 9781119127666.
- William D Callister Jr and David G Rethwisch. *Materials science and engineering: an introduction*. John wiley & sons, 2020.
- Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, et al. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*, 2025. URL <https://arxiv.org/abs/2504.13914>. Accessed: 2025-05-15.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- DeepMind. Ai achieves silver-medal standard solving international mathematical olympiad problems, 2024. URL <https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/>.
- Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.
- Jingxuan Fan, Sarah Martinson, Erik Y Wang, Kaylie Hausknecht, Jonah Brenner, Danxian Liu, Nianli Peng, Corey Wang, and Michael P Brenner. Hardmath: A benchmark dataset for challenging problems in applied mathematics. *arXiv preprint arXiv:2410.09988*, 2024.
- Mahmoud M Farag. *Materials and process selection for engineering design*. CRC press, 2020.
- Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. Physics: Benchmarking foundation models on university-level physics problem solving. *arXiv preprint arXiv:2503.21821*, 2025.

- Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024.
- Joel R Fried. *Polymer science and technology*. Pearson Education, 2014.
- Pradeep Fulay and Jung-Kun Lee. *Electronic, magnetic, and optical materials*. Crc Press, 2016.
- Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*, 2024.
- Google DeepMind. Introducing gemini 2.0: Our new ai model for the agentic era. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message>, December 2024. Accessed: 2025-05-15.
- Google DeepMind. Gemini pro. <https://deepmind.google/technologies/gemini/pro/>, 2025. Accessed: 2025-05-15.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*, 2023.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Ashim Gupta, Vivek Gupta, Shuo Zhang, Yujie He, Ning Zhang, and Shalin Shah. Enhancing question answering on charts through effective pre-training tasks. *arXiv preprint arXiv:2406.10085*, 2024.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- William F Hosford. *Mechanical behavior of materials*. Cambridge university press, 2010a.
- William F Hosford. *Physical metallurgy*. CRC press, 2010b.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.
- Yuqing Huang, Rongyang Zhang, Xuesong He, Xuyang Zhi, Hao Wang, Xin Li, Feiyang Xu, Deguang Liu, Huadong Liang, Yi Li, et al. Chemeval: A comprehensive multi-level chemical evaluation for large language models. *arXiv preprint arXiv:2409.13989*, 2024a.
- Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyuman-shan Ye, Ethan Chern, Yixin Ye, et al. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, 37:19209–19253, 2024b.
- Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36:39648–39677, 2023.
- Jakub Lála, Odhran O’Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodrigues, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

- Chenyue Li, Wen Deng, Mengqian Lu, and Binhang Yuan. Atmoss-ci-bench: Evaluating the recent advance of large language model for atmospheric science. *arXiv preprint arXiv:2502.01159*, 2025a.
- Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. Biomedrag: A retrieval augmented large language model for biomedicine. *Journal of Biomedical Informatics*, 162:104769, 2025b.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.
- Mingyu Derek Ma, Chenchen Ye, Yu Yan, Xiaoxuan Wang, Peipei Ping, Timothy S Chang, and Wei Wang. Clibench: A multifaceted and multigranular evaluation of large language models for clinical decision making. *arXiv preprint arXiv:2406.09923*, 2024a.
- Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, et al. Sciagent: Tool-augmented language models for scientific reasoning. *arXiv preprint arXiv:2402.11451*, 2024b.
- Pragati Shuddhodhan Meshram, Swetha Karthikeyan, Suma Bhat, et al. Electrovizqa: How well do multi-modal llms perform in electronics visual question answering? *arXiv preprint arXiv:2412.00102*, 2024.
- Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, April 2025. Accessed: 2025-05-15.
- Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1527–1536, 2020.
- Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.
- Mistral AI Team. Mistral ocr, March 2025. URL <https://mistral.ai/news/mistral-ocr>. Accessed: 2025-09-22.
- OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, July 2024. Accessed: 2025-05-15.
- OpenAI. Gpt-4.1: The complete guide. <https://gpt-4-1.com>, 2025. Accessed: 2025-05-15.
- OpenAI. Introducing gpt-5. <https://openai.com/index/introducing-gpt-5/>, 2025. Accessed: 2025-09-24.
- OpenAI. Openai o3-mini, 2025. URL <https://openai.com/index/openai-o3-mini/>. Accessed: 2025-05-15.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,

- Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.
- Varuni Sarwal, Seungmo Lee, Rosemary He, Aingela Kattapuram, Eleazar Eskin, Wei Wang, Serghei Mangul, et al. Bioinformaticsbench: A collaboratively built large language model benchmark for bioinformatics reasoning. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*, 2024.
- James F. Shackelford. *Introduction to Materials Science for Engineers*. Pearson, Upper Saddle River, NJ, 8 edition, 2015.
- James F Shackelford. *Introduction to materials science for engineers*. Pearson Upper Saddle River, 2016.
- James F. Shackelford, Young-Hwan Han, Sukyoung Kim, and Se-Hun Kwon. *CRC Materials Science and Engineering Handbook*. CRC Press, Boca Raton, FL, 4 edition, 2016.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAI Conference on Artificial Intelligence*, volume 38, pages 19053–19061, 2024.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petri, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evcı, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris

- Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.
- Daniel Truhn, Jorge S Reis-Filho, and Jakob Nikolas Kather. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nature medicine*, 29(12):2983–2984, 2023.
- George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. Putnambench: A multilingual competition-mathematics benchmark for formal theorem-proving. In *AI for Math Workshop@ ICML 2024*, 2024.
- Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*, 2017.
- Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. <https://arxiv.org/abs/2505.09388>, May 2025. arXiv:2505.09388 [cs.CL].
- Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024a.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b.
- Mohd Zaki, NM Anoop Krishnan, et al. Mascqa: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327, 2024.
- Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaying Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054*, 2025.

## CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Related Work</b>	<b>3</b>
<b>3</b>	<b>Dataset</b>	<b>3</b>
3.1	Data Collection and Processing . . . . .	3
3.2	Dataset Statistics . . . . .	4
3.3	Taxonomy Classification . . . . .	4
3.4	Difficulty Classification . . . . .	5
3.5	Data Leakage Detection . . . . .	5
<b>4</b>	<b>Experiments</b>	<b>6</b>
4.1	Models and Methods . . . . .	6
4.2	Evaluation . . . . .	6
4.3	Results . . . . .	7
4.4	Performance of Small Language Models on <b>MATRIX</b> . . . . .	8
<b>5</b>	<b>Analysis</b>	<b>8</b>
5.1	Performance across Difficulty Levels . . . . .	8
5.2	Efficiency v.s. Accuracy . . . . .	9
5.3	Performance Drop Due to Visual Context . . . . .	9
5.4	Failure Pattern . . . . .	9
5.5	Web Search-based Retrieval Augmented Generation: A Case Study . . . . .	10
<b>6</b>	<b>Conclusion and Limitation</b>	<b>10</b>
<b>A</b>	<b>Additional Details about Dataset</b>	<b>17</b>
A.1	Data Editing UI . . . . .	17
A.2	Taxonomy Tree . . . . .	18
A.3	Comparison of Difficulty Assessment Methods . . . . .	18
A.3.1	Response-Length Based Difficulty Classification . . . . .	18
A.3.2	Step-Count Based Difficulty Classification . . . . .	18
A.4	Example of Questions from Each Difficulty Level . . . . .	20
<b>B</b>	<b>Additional Experiments Details</b>	<b>22</b>
B.1	Details of Different Prompts . . . . .	22
<b>C</b>	<b>Additional Analysis Details</b>	<b>23</b>
C.1	Detailed Performance Across Difficulty Level . . . . .	23
C.2	Details of Error Categorizations . . . . .	23

C.3 RAG Analysis . . . . . 40

A ADDITIONAL DETAILS ABOUT DATASET

A.1 DATA EDITING UI

The user interface of our data editing app is presented in Figure 7. This UI present the QA and allow users to edit each field of the QA.

**Fundamentals\_of\_Materials\_Instructors\_Example 6.37**

**Question Source**

A diffusion couple similar to that shown in Figure 6.1a is prepared using two hypothetical metals A and B. After a 20-h heat treatment at 800°C (and subsequently cooling to room temperature) the concentration of B is 5.0 wt% at the 5.0-mm position within metal A. If another heat treatment is conducted on an identical diffusion couple, but at 1000°C for 20 h, at what position will the composition be 5.0 wt% B? Assume that the preexponential and activation energy for the diffusion coefficient are  $1.5 \times 10^{-4} \text{ m}^2/\text{s}$  and  $125,000 \text{ J/mol}$ , respectively.

**Question Preview**

A diffusion couple similar to that shown in Figure 6.1a is prepared using two hypothetical metals A and B. After a 20-h heat treatment at 800°C (and subsequently cooling to room temperature) the concentration of B is 5.0 wt% at the 5.0-mm position within metal A. If another heat treatment is conducted on an identical diffusion couple, but at 1000°C for 20 h, at what position will the composition be 5.0 wt% B? Assume that the preexponential and activation energy for the diffusion coefficient are  $1.5 \times 10^{-4} \text{ m}^2/\text{s}$  and  $125,000 \text{ J/mol}$ , respectively.

**Solution Source**

In order to determine the position within the diffusion couple at which the concentration of A in B is 5.0 wt%, we must employ equation (D-2) (with  $D_0 = 1.5 \times 10^{-4} \text{ m}^2/\text{s}$  and  $Q_d = 125,000 \text{ J/mol}$ ). That is

$$C_B = C_0 \exp\left(-\frac{D_0}{D} \exp\left(-\frac{Q_d}{RT}\right)\right)$$

or

$$\frac{D}{D_0} = \frac{1}{C_0} \exp\left(-\frac{Q_d}{RT}\right) \exp\left(-\frac{C_B}{C_0}\right)$$

It is first necessary to compute values for both  $D_{800}$  and  $D_{1000}$ ; this is accomplished using equation (D-1) (with  $D_0 = 1.5 \times 10^{-4} \text{ m}^2/\text{s}$  and  $Q_d = 125,000 \text{ J/mol}$ ) as follows:

$$D_{800} = (1.5 \times 10^{-4} \text{ m}^2/\text{s}) \exp\left(-\frac{125,000 \text{ J/mol}}{(8.31 \text{ J/mol}\cdot\text{K})(800 + 273 \text{ K})}\right) = 1.22 \times 10^{-10} \text{ m}^2/\text{s}$$

$$D_{1000} = (1.5 \times 10^{-4} \text{ m}^2/\text{s}) \exp\left(-\frac{125,000 \text{ J/mol}}{(8.31 \text{ J/mol}\cdot\text{K})(1000 + 273 \text{ K})}\right) = 1.11 \times 10^{-9} \text{ m}^2/\text{s}$$

Now, solving equation (D-2) (with  $C_B = 5.0$  wt% and  $C_0 = 100$  wt%) yields

$$x = \sqrt{\frac{D_{1000}}{D_{800}}} \sqrt{\frac{1.11 \times 10^{-9} \text{ m}^2/\text{s}}{1.22 \times 10^{-10} \text{ m}^2/\text{s}}} = 15.1 \text{ mm}$$

**Answer Source**

15.1

**Answer Preview**

15.1

**Unit**

mm

**Unit Preview**

mm

**Categories**

Materials	Properties	Structures
Metals	Thermal	Composites
Fundamental Mechanisms	Processes	Failure Mechanisms

**Type**

MCQ

**Number of Answers**

Steps: 1

Save & Reformat

Previous Next [Submit Answer](#)

**Images**

**Figure 6.1** (a) A copper-nickel diffusion couple before a high-temperature heat treatment. (b) Schematic representations of Cu (red circles) and Ni (blue circles) atom locations within the diffusion couple. (c) Concentrations of copper and nickel as a function of position across the couple.

**Upload New Image**

Choose File. No file selected.

**Notes**

**Notes Preview**

Figure 7: UI of Data Editing App

## A.2 TAXONOMY TREE

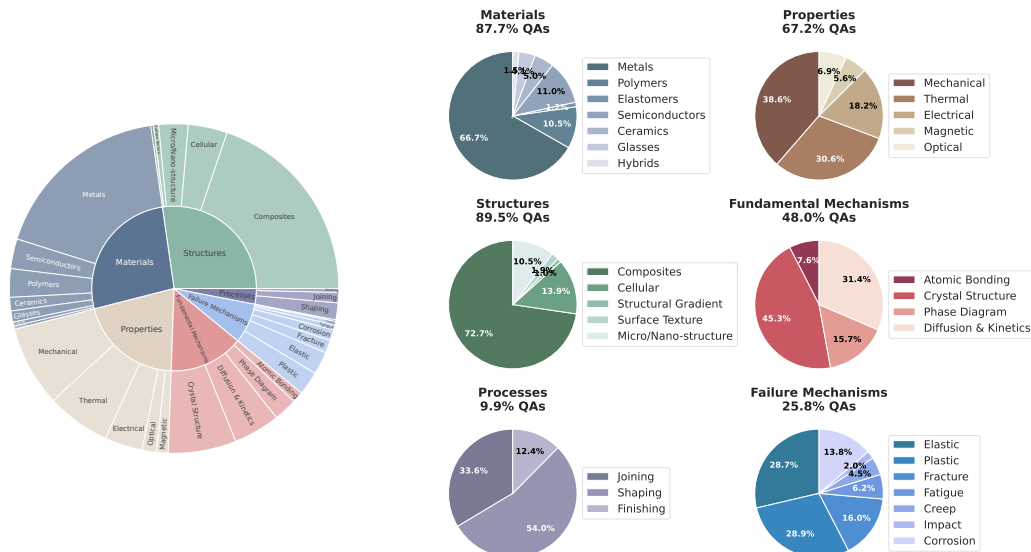


Figure 8: Taxonomy and QA distribution across each fields and sub-fields.

The proposed taxonomy for materials science QAs includes 6 major fields. For each question, we assign a six-digit “category vector” where each position corresponds to one of these primary fields. The digit in each position indicates the relevant subcategory within that field, with 0 representing “None” (indicating the field is not relevant to the question). For example, a question with category vector “310001” would be about Metals, Mechanical properties, with no relevant Structures, Fundamental Mechanisms, or Processes components, and related to Elastic deformation. Only 10 questions are not matched with any subfield and are manually assigned to “Mechanical Properties” (“010000”).

Figure 8 shows the distribution across taxonomy subcategories. This multi-dimensional classification scheme enables us to capture the interdisciplinary nature of materials science problems, where a single question might span multiple domains. Only 10 questions are not matched with any subfield and are manually assigned to “Mechanical Properties” (“010000”).

## A.3 COMPARISON OF DIFFICULTY ASSESSMENT METHODS

### A.3.1 RESPONSE-LENGTH BASED DIFFICULTY CLASSIFICATION

We evaluate question difficulty based on response lengths from a strong baseline model, Claude-3.7-Sonnet. The questions are classified into three tiers: Easy, Medium, and Hard. We applied this classification to other models to demonstrate its robustness. As shown in Table 6, for nearly every model, accuracy decreases while response length (characters) increases with difficulty, confirming the reliability of this assessment method.

The consistent pattern across multiple models validates this classification methodology, showing that longer required responses correspond to more complex reasoning requirements. We adopt this response-length based classification as our primary difficulty assessment method, with step-count analysis serving as validation.

### A.3.2 STEP-COUNT BASED DIFFICULTY CLASSIFICATION

To ensure robustness in our difficulty classification, we employed three complementary approaches to determine solution step counts:

- 1. Pattern-based method:** This approach identifies explicit step indicators in solution text, such as numbered steps, paragraph breaks, and calculation indicators.

Table 6: Model Performance and Response Length across Difficulty Levels Based on Response-Length Classification

Model	Accuracy			Response Length		
	Easy	Medium	Hard	Easy	Medium	Hard
claude-3-7-sonnet	0.640	0.541	0.480	1180	1496	1947
deepseek-V3	0.712	0.558	0.487	1413	1690	2310
deepseek-reasoner	0.749	0.658	0.592	15901	18417	20689
gemini-2.0-flash	0.552	0.484	0.363	1713	2316	2816
gemini-2.5-pro-preview	0.632	0.551	0.527	3209	3593	4331
gpt-4.1	0.591	0.506	0.414	1382	1786	2300
llama-4-maverick	0.693	0.553	0.495	2157	2626	3096
o4-mini	0.586	0.471	0.498	400	467	505
qwen3-235b-a22b	0.745	0.657	0.563	1640	1917	2188

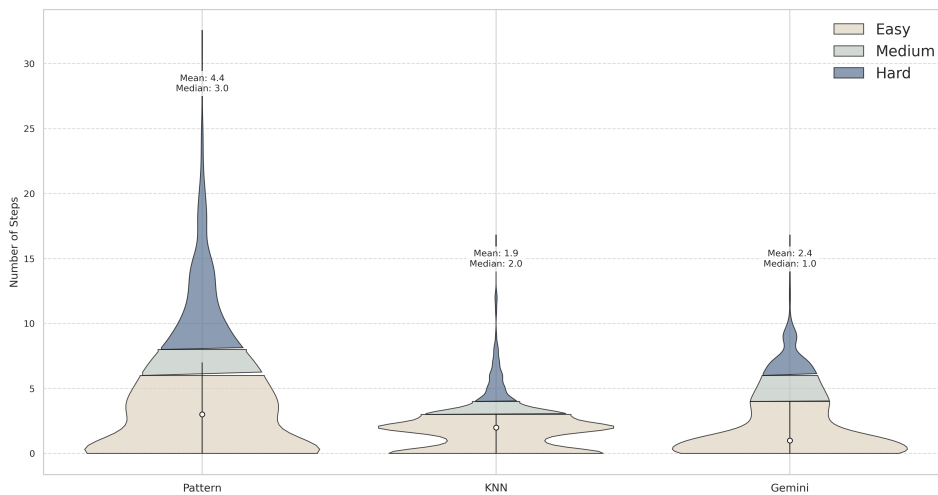


Figure 9: Distribution of solution steps count across the dataset.

2. **KNN-based method:** Trained on 100 sampled examples, this approach uses few-shot learning to identify implicit solution steps even when they aren't explicitly numbered.
3. **Gemini-based method:** This method leverages the Gemini-2.0-flash model to analyze solution structure and identify both explicit and implicit reasoning steps. It produces a mean of 2.4 steps with a median of 1.0 steps, resulting in our final difficulty distribution.

As illustrated in Figure 9, the distributions of solution steps across the three strategies follow a right-skewed pattern, with most questions requiring fewer than 4 steps to solve. While pattern-based method potentially overestimates the steps needed.

These three approaches show moderate agreement in their classifications, with pairwise agreement rates of 57.1% between Pattern-based and Gemini-based methods, 47.9% between Pattern-based and KNN-based methods, and 44.2% between KNN-based and Gemini-based methods. We analyzed the correlation of step counts between the methods, which reveals that Pattern vs. Gemini has high correlation (0.61), confirming that our Gemini-based assessment captures many of the explicit steps identified by pattern matching while also recognizing implicit reasoning steps.

We further verified the correlation of solution step number from the three counting strategy in Figure 10 and the agreement of difficulty level derived from three step counting methods in Figure 11.

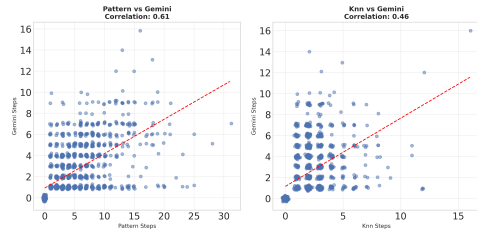


Figure 10: Correlation analysis of three step counting methods.

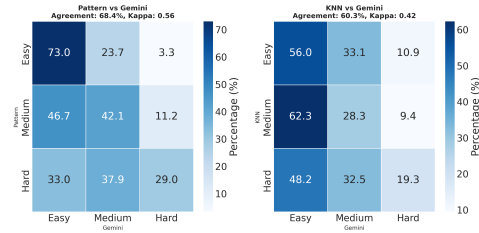


Figure 11: Agreement of difficulty level derived from three step counting methods.

#### A.4 EXAMPLE OF QUESTIONS FROM EACH DIFFICULTY LEVEL

Here we present questions from each difficulty level.

##### Example: Easy

###### Question

If ice homogeneously nucleates at  $-40^\circ\text{C}$ , calculate the critical radius given values of  $-3.1 \times 10^8 \text{ J/m}^3$  and  $25 \times 10^{-3} \text{ J/m}^2$ , respectively, for the latent heat of fusion and the surface free energy.

###### Solution

This problem states that ice homogeneously nucleates at  $-40^\circ\text{C}$ , and we are to calculate the critical radius given the latent heat of fusion ( $-3.1 \times 10^8 \text{ J/m}^3$ ) and the surface free energy ( $25 \times 10^{-3} \text{ J/m}^2$ ). Solution to this problem requires the utilization of equation

$$r^* = \left( -\frac{2\gamma T_m}{\Delta H_f} \right) \left( \frac{1}{T_m - T} \right)$$

as

$$\begin{aligned} r &= \left( -\frac{2\gamma T_m}{\Delta H_f} \right) \left( \frac{1}{T_m - T} \right) \\ &= \left[ -\frac{(2)(25 \times 10^{-3} \text{ J/m}^2)(273 \text{ K})}{-3.1 \times 10^8 \text{ J/m}^3} \right] \left( \frac{1}{40 \text{ K}} \right) \\ &= 1.10 \times 10^{-9} \text{ m} = 1.10 \text{ nm} \end{aligned}$$

##### Example: Medium

###### Question

You are asked to characterize a new semiconductor. If its conductivity at  $20^\circ\text{C}$  is  $250\Omega^{-1} \cdot \text{m}^{-1}$  and at  $100^\circ\text{C}$  is  $1,100\Omega^{-1} \cdot \text{m}^{-1}$ , what is its band gap,  $E_g$ ?

###### Solution

From Equation ( $\ln \sigma = \ln \sigma_0 - \frac{E_g}{2k} \cdot \frac{1}{T}$ ),

$$\ln \sigma_{T_1} = \ln \sigma_0 - \frac{E_g}{2k} \cdot \frac{1}{T_1}$$

and

$$\ln \sigma_{T_2} = \ln \sigma_0 - \frac{E_g}{2k} \cdot \frac{1}{T_2}$$

Subtracting the second from the first yields:

$$\begin{aligned}\ln \sigma_{T_1} - \ln \sigma_{T_2} &= \ln \left( \frac{\sigma_{T_1}}{\sigma_{T_2}} \right) \\ &= -\frac{E_g}{2k} \left( \frac{1}{T_1} - \frac{1}{T_2} \right)\end{aligned}$$

Then,

$$-\frac{E_g}{2k} = \frac{\ln(\sigma_{T_1}/\sigma_{T_2})}{1/T_1 - 1/T_2}$$

or

$$E_g = \frac{2k \cdot \ln(\sigma_{T_2}/\sigma_{T_1})}{1/T_1 - 1/T_2}$$

Taking  $T_1 = 20^\circ\text{C} (= 293 \text{ K})$  and  $T_2 = 100^\circ\text{C} (= 373 \text{ K})$  gives:

$$\begin{aligned}E_g &= \frac{(2 \times 86.2 \times 10^{-6} \text{ eV/K}) \cdot \ln(1,100/250)}{\frac{1}{373} - \frac{1}{293}} \\ &= 0.349 \text{ eV}\end{aligned}$$

### Example: Hard

#### Question

An advanced engineered ceramic has a Weibull modulus  $m = 9$ . The flexural strength is 250 MPa at a probability of failure  $F = 0.4$ . What is the level of flexural strength if the probability of failure has to be 0.1?

#### Solution

We assume all samples tested had the same volume, so the size of the sample will not be a factor in this case. We can use the symbol  $V$  for sample volume instead of  $V_0$ . We are dealing with a brittle material, so we begin with the equation:

$$F(V) = 1 - P(V) = 1 - \exp \left[ - \left( \frac{\sigma}{\sigma_0} \right)^m \right]$$

or

$$1 - F(V) = \exp \left[ - \left( \frac{\sigma}{\sigma_0} \right)^m \right]$$

Take the logarithm of both sides:

$$\ln[1 - F(V)] = - \left( \frac{\sigma}{\sigma_0} \right)^m$$

Take logarithms again:

$$\ln \{ -\ln[1 - F(V)] \} = m (\ln \sigma - \ln \sigma_0)$$

We eliminate the minus sign on the left-hand side by rewriting as:

$$\ln \left\{ \ln \left[ \frac{1}{1 - F(V)} \right] \right\} = m (\ln \sigma - \ln \sigma_0)$$

For  $F = 0.4$ ,  $\sigma = 250 \text{ MPa}$ , and  $m = 9$ , we have:

$$\ln \left[ \ln \left( \frac{1}{1 - 0.4} \right) \right] = 9 (\ln 250 - \ln \sigma_0)$$

Therefore,

$$\begin{aligned}\ln[\ln(1/0.6)] &= \ln[\ln(1.66667)] = \ln(0.510826) = -0.67173 \\ &= 9(5.52146 - \ln \sigma_0)\end{aligned}$$

Solving gives:

$$\ln \sigma_0 = 5.52146 + 0.07464 = 5.5961 \Rightarrow \sigma_0 = 269.4 \text{ MPa}$$

Now, to find the value of  $\sigma$  for  $F = 0.1$ , we use the same equation:

$$\begin{aligned}\ln \left[ \ln \left( \frac{1}{1-0.1} \right) \right] &= 9(\ln \sigma - \ln 269.4) \\ \ln[\ln(1/0.9)] &= 9(\ln \sigma - 5.5962) \\ \ln(0.105361) &= -2.25037 = 9(\ln \sigma - 5.5962) \\ &\Rightarrow \ln \sigma = 5.3462\end{aligned}$$

So,

$$\sigma = 209.8 \text{ MPa}$$

As expected, when the probability of failure is reduced to 0.1, the stress level that can be supported also decreases.

## B ADDITIONAL EXPERIMENTS DETAILS

### B.1 DETAILS OF DIFFERENT PROMPTS

Prompts we used for each method are as follows. The Basic System Prompt is used in basic CoT, self-correction, and the RAG.

#### Basic System Prompt

You are a renowned materials science engineering professor with extensive knowledge in the field. Your students have presented you with a challenging question related to materials science. Please reason step by step, and put the final answer inside a single box using `\boxed{...}`. Include only the final answer inside the box, without the unit.

The tool-augmentation is prompted to use Python code to improve the computation.

#### Tool System Prompt

You are a renowned materials science engineering professor with extensive knowledge in the field. Your students have presented you with a challenging question related to materials science. If necessary, you could write a single clean Python code block that computes necessary numeric values. Enclose the code in triple backticks with `"`python`". Please reason step by step, if no code is needed, put the final answer inside a single box using \boxed{...}; otherwise, wait for the user to execute the code and give you the execution result, and then put the final answer inside a single box using \boxed{...}. Include only the final answer inside the box, without the unit.`

After the code execution, the model get the results and make the final answer.

#### Tool Summary Prompt

Here are the results of the code execution:  
`\n\n{code_executed}\n\n`Based on these results, what is the final answer to the original question?

When using the self-correction, the model is first prompted to review and find the problem from its initial response.

**Review Prompt**

Review your previous answer and find problems with your answer.

Then, the model is prompted to improve the initial response with the problem it found.

**Revise Prompt**

Based on the problems you found, improve your answer. Please reiterate your answer, with your final answer in the form `\boxed{answer}`.

We use the following prompt to let Gemini-2.0-Flash determine whether the answer is correct.

**Judge System Prompt**

As an expert judge, evaluate if the following model’s answer matches the reference answer.\n Focus on the numerical values and key concepts. Small numerical differences are tolerable due to approximation errors.\n Do not solve the problem, just judge if the model answer matches the reference answer.\n Put the final decision ('correct' (if matching) or 'incorrect' (if not matching)) inside a single box using `\boxed{...}`.

**C ADDITIONAL ANALYSIS DETAILS****C.1 DETAILED PERFORMANCE ACROSS DIFFICULTY LEVEL**

The performance of each model across difficulty levels is presented in Table 7.

**C.2 DETAILS OF ERROR CATEGORIZATIONS**

We use Gemini-2.0-Flash to categorize the error using the following prompt:

**Error Categorization Prompt**

You are an assistant whose task is to diagnose the single main reason a wrong solution fails. Each task input will contain three parts, clearly marked: (i) the question, (ii) a reference solution, and (iii) a wrong solution produced by a model. Your steps are:

1. Read the question first so you know what must be answered. Pay attention to given data, required units, and any boundary conditions or hidden assumptions.
2. Read the reference solution carefully. Treat it as correct and complete unless it contains an explicit note that it is partial.
3. Read the wrong solution line by line. Locate the first point where it diverges from the reasoning in the reference solution. That first wrong turn usually signals the true cause of failure.

Choose one category below that best explains the root cause. If more than one category is possible, pick the one that triggers the earliest error or has the largest impact on the final answer. If the wrong solution actually reaches the same numerical result and its reasoning is valid, assign category 7.

**Categories**

1. Problem Comprehension and Assumptions. The solver misreads what is asked, drops a given fact, injects an unsupported assumption, or confuses symbols.
2. Domain Knowledge Accuracy. The solver recalls or applies a materials science law, concept, or formula in an incorrect way. Unit definitions and physical constants also belong here when misused.
3. Solution Strategy and Planning. The solver sets up an approach that cannot reach the goal, skips required sub-problems, or mixes independent lines of reasoning without a clear plan.

Table 7: Experimental Results by Difficulty Level on **MATRIX** (questions w/o images). **Bold** indicates the best performance, and Underline indicates the second best.

Model	Easy	Medium	Hard	Overall
<b>Llama-4-Maverick-17b</b>	79.72	68.09	57.28	71.61
+Correction	78.94	64.80	55.87	69.95
+Tool	75.79	62.83	57.75	68.20
<b>Claude-3.7-Sonnet</b>	75.98	63.49	52.11	67.32
+Correction	77.76	61.84	53.52	68.00
+Tool	78.54	68.09	59.62	71.51
<b>GPT-4.1</b>	76.97	68.75	58.69	70.73
+Correction	75.00	64.47	56.34	68.00
+Tool	70.28	56.91	47.89	61.66
<b>DeepSeek-V3</b>	75.59	60.86	51.17	66.15
+Correction	73.62	57.57	52.11	64.39
+Tool	70.08	59.87	47.89	62.44
<b>Gemini-2.0-Flash</b>	68.11	59.21	41.31	59.90
+Correction	69.88	60.53	46.95	62.34
+Tool	78.74	66.12	52.11	69.46
<b>Qwen2.5-7b</b>	26.38	14.47	6.57	18.73
+Correction	27.36	13.16	8.45	19.22
+Tool	18.31	11.84	2.82	13.17
<b>Qwen2.5-32b</b>	42.13	29.61	17.37	33.27
+Correction	41.34	27.30	16.90	32.10
+Tool	16.14	13.82	8.45	13.85
<b>Gemma-3-4b</b>	9.65	5.92	2.82	7.12
+Correction	7.48	5.92	2.82	6.05
+Tool	8.46	5.92	3.76	6.73
<b>Gemma-3-12b</b>	19.88	9.87	4.69	13.76
+Correction	20.67	12.83	4.23	14.93
+Tool	22.83	12.50	7.51	16.59
<b>Gemma-3-27b</b>	37.60	30.92	15.02	30.93
+Correction	33.46	29.28	16.90	28.78
+Tool	31.10	25.33	15.96	26.24
<b>Gemini-2.5-Pro</b>	<b>82.28</b>	<b>74.67</b>	<b>69.48</b>	<b>77.37</b>
<b>o4-mini</b>	<u>79.13</u>	68.42	<u>71.36</u>	<u>74.34</u>
<b>DeepSeek-Reasoner</b>	77.95	<u>72.37</u>	66.67	73.95
<b>Qwen3-235B</b>	76.38	71.05	63.38	72.10
<b>GPT-5</b>	71.06	63.49	52.11	64.88
<b>Claude-Sonnet-4</b>	60.43	53.62	41.31	54.44

4. Calculation Accuracy. The algebra, arithmetic, sign handling, or unit conversion is wrong even though the plan and formulae are correct.
5. Hallucinated Content. The solver invents inputs, processes, or physical relations that are not stated in the question and are not accepted scientific facts.
6. Code Implementation. The solver writes Python code that does not match its verbal reasoning or has syntax, logic, or data handling errors that change the outcome.
7. Other. Any issue not covered above, or the wrong solution is actually correct.

**Answer format**

Return exactly one  $\text{\TeX}$  box with the chosen index:  $\text{\boxed{1}}$ ,  $\text{\boxed{2}}$ ,  $\text{\boxed{3}}$ ,  $\text{\boxed{4}}$ ,  $\text{\boxed{5}}$ ,  $\text{\boxed{6}}$ , or  $\text{\boxed{7}}$ . Output nothing else.

Here are examples of each category:

**Example: Problem Comprehension and Assumptions****Question:**

Compute the rate of some reaction that obeys Avrami kinetics, assuming that the constants  $n$  and  $k$  have values of 2.0 and  $5 \times 10^{-4}$ , respectively, for time expressed in seconds. The unit of the answer is  $\text{s}^{-1}$ .

**Reference Solution:**

This problem asks that we compute the rate of some reaction given the values of  $n$  and  $k$  in equation

$$(y = 1 - \exp(-kt^n))$$

. Since the reaction rate is defined by equation ( $\text{rate} = \frac{1}{t_{0.5}}$ ), it is first necessary to determine  $t_{0.5}$ , or the time necessary for the reaction to reach  $y = 0.5$ . We must manipulate equation ( $y = 1 - \exp(-kt^n)$ ) such that  $t$  is the dependent variable. It is first necessary to rearrange equation ( $y = 1 - \exp(-kt^n)$ ) as follows:

$$\exp(-kt^n) = 1 - y$$

And then take natural logarithms of both sides:

$$-kt^n = \ln(1 - y)$$

Which may be rearranged so as to read

$$t^n = -\frac{\ln(1 - y)}{k}$$

Now, solving for  $t$  from this expression leads to

$$t = \left[ -\frac{\ln(1 - y)}{k} \right]^{\frac{1}{n}}$$

For  $t_{0.5}$  this equation takes the form

$$t_{0.5} = \left[ -\frac{\ln(1 - 0.5)}{k} \right]^{\frac{1}{n}}$$

And, incorporation of values for  $n$  and  $k$  given in the problem statement ( $2.0$  and  $5 \times 10^{-4}$ , respectively), then

$$t_{0.5} = \left[ -\frac{\ln(1 - 0.5)}{5 \times 10^{-4}} \right]^{\frac{1}{2}} = 37.23 \text{ s}$$

Now, the rate is computed using equation ( $\text{rate} = \frac{1}{t_{0.5}}$ ) as

$$\text{rate} = \frac{1}{t_{0.5}} = \frac{1}{37.23 \text{ s}} = 2.69 \times 10^{-2} \text{ s}^{-1}$$

**DeepSeek-V3 Answer:** To compute the reaction rate  $\frac{dy}{dt}$  for Avrami kinetics with  $n = 2.0$  and  $k = 5 \times 10^{-4} \text{ s}^{-n}$ , we use the formula:

$$\frac{dy}{dt} = nkt^{n-1} \exp(-kt^n)$$

Substituting the given values:

$$\frac{dy}{dt} = 2.0 \times 5 \times 10^{-4} \times t^{2-1} \exp(-5 \times 10^{-4} \times t^2)$$

Simplifying:

$$\frac{dy}{dt} = 10^{-3}t \exp(-5 \times 10^{-4}t^2)$$

The reaction rate at any time  $t$  is therefore:

$$10^{-3}t \exp(-5 \times 10^{-4}t^2)$$

**\*\*Note:\*\*** The answer is expressed in terms of  $t$ . If a specific time  $t$  is provided, substitute it into the expression to get the numerical rate. The unit  $\text{s}^{-1}$  is implied.

**Explanation:**

The task assumes the reaction rate as  $\text{rate} = \frac{1}{t_{0.5}}$ , that is, the reciprocal of the time required to reach a fractional conversion of  $y = 0.5$ . The model instead differentiated the Avrami equation to obtain the instantaneous rate  $\frac{dy}{dt}$  as a function of time, left the result in symbolic form, and never evaluated it at the half-transformation point or computed a numerical value.

### Domain Knowledge Accuracy

**Question**

Given  $r_{\text{Si}} = 0.117 \text{ nm}$ , calculate the linear density of atoms along the [111] direction in silicon. Express the answer in  $\frac{\text{atoms}}{\text{nm}}$ .

**Reference Solution**

To calculate the linear density of atoms along the [111] direction in silicon, we must consider the arrangement of atoms along this direction, which is a body diagonal in the diamond cubic structure. The atoms along the [111] direction are not uniformly spaced, and thus, the calculation requires careful consideration of their positions. Referring to the structure, we find that there are effectively 2 atoms centered along a given body diagonal. The length of the body diagonal in a unit cell, denoted as  $l$ , can be related to the atomic radius of silicon,  $r_{\text{Si}}$ , by the equation:

$$2r_{\text{Si}} = \frac{1}{4}l$$

Solving for  $l$  gives:

$$l = 8r_{\text{Si}}$$

. The atomic radius of silicon is  $0.117 \text{ nm}$ , thus:

$$l = 8(0.117 \text{ nm}) = 0.936 \text{ nm}$$

Therefore, the linear density is calculated as:

$$\text{linear density} = \frac{2 \text{ atoms}}{0.936 \text{ nm}} = 2.14 \frac{\text{atoms}}{\text{nm}}$$

**DeepSeek-V3 Answer**

To calculate the linear density of atoms along the [111] direction in silicon, follow these steps:

1. **Determine the lattice parameter  $a$ :**

$$a = \frac{4r}{\sqrt{3}} = \frac{4 \times 0.117 \text{ nm}}{\sqrt{3}} \approx 0.270 \text{ nm}$$

2. **Calculate the distance between atoms along the [111] direction:** The [111] direction in a diamond cubic structure has atoms spaced at intervals of  $\frac{\sqrt{3}}{4}a$ . Therefore:

$$\text{Distance} = \frac{\sqrt{3}}{4}a = \frac{\sqrt{3}}{4} \times 0.270 \text{ nm} \approx 0.117 \text{ nm}$$

3. **Compute the linear density:** The linear density is the inverse of the distance between atoms:

$$\text{Linear Density} = \frac{1}{\text{Distance}} = \frac{1}{0.117 \text{ nm}} \approx 8.55 \frac{\text{atoms}}{\text{nm}}$$

However, upon closer inspection, the [111] direction in the diamond cubic structure actually contains **2 atoms** over a length of  $\frac{\sqrt{3}}{2}a$ . Therefore, the correct linear density is:

$$\text{Linear Density} = \frac{2}{\frac{\sqrt{3}}{2}a} = \frac{4}{\sqrt{3}a}$$

Substituting  $a = 0.270 \text{ nm}$ :

$$\text{Linear Density} = \frac{4}{\sqrt{3} \times 0.270} \approx 8.55 \frac{\text{atoms}}{\text{nm}}$$

Thus, the final answer is:

8.55

### Explanation

The model's result is wrong because they chose the wrong crystallographic relation between atomic radius and lattice parameter for a diamond-cubic lattice: they wrote

$$a = \frac{4r}{\sqrt{3}}$$

instead of the correct

$$a = \frac{8r}{\sqrt{3}}$$

That choice is a mis-remembered structural fact about how silicon atoms touch along  $\langle 110 \rangle$  directions.

### Example: Calculation Accuracy

#### Question:

Equation  $w = \frac{G\gamma^2}{2} = \left(\frac{1}{2}\right) \frac{Gb^2}{(2\pi r)^2}$  gives the energy/volume as a function of  $r$ . Calculate the distance from the core of a screw dislocation at which the energy/volume equals the heat of vaporization,  $H_v$ , and express this distance in terms of  $r/b$ . Evaluate this critical value of  $r/b$  for copper. Given  $\rho = 8.93 \text{ mg/cm}^3$ ,  $G = 77 \text{ GPa}$ ,  $b = 0.255 \text{ nm}$ ,  $H_v = 4.73 \text{ MJ/kg}$ . Give your answer as a tuple: (critical  $r/b$ , critical  $r$  of copper). The units of each required answer are ( $r/b$ , nm), respectively.

#### Correct Solution:

Solving the equation

$$w = \frac{G\gamma^2}{2} = \left(\frac{1}{2}\right) \frac{Gb^2}{(2\pi r)^2}$$

for  $\frac{r}{b}$ , we obtain:

$$\left(\frac{r}{b}\right)^2 = \frac{G}{8\pi^2 U_v} = \frac{77 \times 10^9 \text{ Pa}}{8\pi^2 \times 8.93 \times 10^3 \text{ kg/m}^3 \times 4.73 \times 10^6 \text{ J/kg}} = 0.23,$$

so

$$\frac{r}{b} = 0.15.$$

For copper,

$$r = 0.15 \times 0.255 = 0.04 \text{ nm}.$$

**DeepSeek-V3 Answer:**

1. Equate energy per volume to heat of vaporization:

$$\frac{1}{2} \frac{Gb^2}{(2\pi r)^2} = H_v \Rightarrow \frac{Gb^2}{8\pi^2 r^2} = H_v$$

2. Solve for  $r$ :

$$r^2 = \frac{Gb^2}{8\pi^2 H_v}, \quad r = \frac{b}{2\pi} \sqrt{\frac{G}{2H_v}}$$

3. Express  $r$  in terms of  $r/b$ :

$$\frac{r}{b} = \frac{1}{2\pi} \sqrt{\frac{G}{2H_v}}$$

4. Convert  $H_v$  from MJ/kg to J/m<sup>3</sup>:

$$\rho = 8.93 \text{ g/cm}^3 = 8930 \text{ kg/m}^3, \quad H_v = 4.73 \times 10^6 \text{ J/kg}$$

$$U_v = H_v \times \rho = 4.22 \times 10^{10} \text{ J/m}^3$$

5. Substitute the values:

$$\frac{r}{b} = \frac{1}{2\pi} \sqrt{\frac{77 \times 10^9}{2 \times 4.22 \times 10^{10}}} \approx \frac{1}{2\pi} \sqrt{\frac{77}{8.44}} \approx \frac{1}{2\pi} \times 3.02 \approx 0.48$$

6. Calculate critical  $r$  for copper:

$$r = 0.48 \times 0.255 \text{ nm} \approx 0.122 \text{ nm}$$

**Explanation:**

The model wrote exactly the same physical relation as the reference and kept the density when converting the heat of vaporisation to  $U_v$ . The only thing that went wrong is the arithmetic:

$$\frac{2U_v}{G} = \frac{8.44 \times 10^{10}}{77 \times 10^9} \approx 0.912,$$

but the model treated it as

$$\frac{77}{8.44} \approx 9.13,$$

ignoring the difference in the powers of ten. Taking the square root and then multiplying by  $\frac{1}{2\pi}$  turned 0.15 into 0.48, which in turn made  $r$  three times too large.

**Example: Solution Strategy and Planning**

**Question:**

Copper (Cu) and platinum (Pt) both have the FCC crystal structure, and Cu forms a substitutional solid solution for concentrations up to approximately 6 wt% Cu at room temperature. Determine the concentration in weight percent of Cu that must be added to Pt to yield a unit cell edge length of 0.390 nm. The atomic weights for Cu and Pt are 63.55 and 195.08 g/mol, respectively. Unit of the answer: wt%.

**Reference Solution:**

To begin, it is necessary to employ the equation

$$\rho = \frac{nA}{V_C N_A},$$

and solve for the unit cell volume,  $V_C$ , as

$$V_C = \frac{nA_{\text{ave}}}{\rho_{\text{ave}} N_A},$$

where  $A_{\text{ave}}$  and  $\rho_{\text{ave}}$  are the atomic weight and density, respectively, of the Pt – Cu alloy. Inasmuch as both of these materials have the FCC crystal structure, which has cubic symmetry,  $V_C$  is just the cube of the unit cell length,  $a$ . That is,

$$\begin{aligned} V_C &= a^3 = (0.390 \text{ nm})^3 \\ &= (3.90 \times 10^{-8} \text{ cm})^3 = 5.932 \times 10^{-23} \text{ cm}^3 \end{aligned}$$

It is now necessary to construct expressions for  $A_{\text{ave}}$  and  $\rho_{\text{ave}}$  in terms of the concentration of copper,  $C_{\text{Cu}}$ , using the equations

$$\rho_{\text{ave}} = \frac{100}{\frac{C_1}{\rho_1} + \frac{C_2}{\rho_2}}, \quad A_{\text{ave}} = \frac{100}{\frac{C_1}{A_1} + \frac{C_2}{A_2}}.$$

For  $A_{\text{ave}}$ , we have

$$\begin{aligned} A_{\text{ave}} &= \frac{100}{\frac{C_{\text{Cu}}}{A_{\text{Cu}}} + \frac{100 - C_{\text{Cu}}}{A_{\text{Pt}}}} \\ &= \frac{100}{\frac{C_{\text{Cu}}}{63.55 \text{ g/mol}} + \frac{100 - C_{\text{Cu}}}{195.08 \text{ g/mol}}} \end{aligned}$$

since the atomic weights for Cu and Pt are, respectively, 63.55 and 195.08 g/mol. Now, the expression for  $\rho_{\text{ave}}$  is as follows:

$$\begin{aligned} \rho_{\text{ave}} &= \frac{100}{\frac{C_{\text{Cu}}}{\rho_{\text{Cu}}} + \frac{100 - C_{\text{Cu}}}{\rho_{\text{Pt}}}} \\ &= \frac{100}{\frac{C_{\text{Cu}}}{8.94 \text{ g/cm}^3} + \frac{100 - C_{\text{Cu}}}{21.45 \text{ g/cm}^3}} \end{aligned}$$

given the densities of 8.94 and 21.45 g/cm<sup>3</sup> for the respective metals. Within the FCC unit cell, there are 4 equivalent atoms, and thus, the value of  $n$  in the equation

$$\rho = \frac{nA}{V_C N_A}$$

is 4; hence, the expression for  $V_C$  may be written in terms of the concentration of Cu in weight percent as follows:

$$\begin{aligned} V_C &= 5.932 \times 10^{-23} \text{ cm}^3 \\ &= \frac{nA_{\text{ave}}}{\rho_{\text{ave}} N_A} \end{aligned}$$

And solving this expression for  $C_{\text{Cu}}$  leads to

$$C_{\text{Cu}} = 2.83 \text{ wt}\%.$$

**DeepSeek-V3 Answer**

To determine the weight percent of Cu that must be added to Pt to achieve a unit cell edge length of 0.390 nm in an FCC structure, follow these steps:

**\*\*Given:\*\*** - FCC unit cell edge length ( $a$ ) = 0.390 nm =  $0.390 \times 10^{-7}$  cm - Atomic weights:  
 - Cu ( $A_{\text{Cu}}$ ) = 63.55 g/mol - Pt ( $A_{\text{Pt}}$ ) = 195.08 g/mol - FCC structure has 4 atoms per unit cell.  
**\*\*Step 1:** Calculate the unit cell volume ( $V_{\text{cell}}$ )\*\*

$$V_{\text{cell}} = a^3 = (0.390 \times 10^{-7} \text{ cm})^3 = 5.9319 \times 10^{-23} \text{ cm}^3$$

**\*\*Step 2:** Relate the unit cell volume to the average atomic volume\*\* For an FCC structure, the volume per atom ( $V_{\text{atom}}$ ) is:

$$V_{\text{atom}} = \frac{V_{\text{cell}}}{4} = \frac{5.9319 \times 10^{-23}}{4} = 1.4830 \times 10^{-23} \text{ cm}^3$$

The average atomic volume is also related to the atomic weights and densities. However, we can express the average atomic radius ( $r$ ) in terms of the unit cell edge length for FCC:

$$a = 2\sqrt{2}r \implies r = \frac{a}{2\sqrt{2}} = \frac{0.390}{2\sqrt{2}} = 0.1379 \text{ nm}$$

But instead, we use the **\*\*average atomic volume\*\*** approach to find the **\*\*average atomic weight ( $A_{\text{avg}}$ )\*\***.

**\*\*Step 3:** Relate the average atomic volume to the average atomic weight\*\* The volume per atom is also given by:

$$V_{\text{atom}} = \frac{A_{\text{avg}}}{\rho_{\text{avg}} N_A}$$

where: -  $\rho_{\text{avg}}$  = average density -  $N_A$  = Avogadro's number =  $6.022 \times 10^{23}$  atoms/mol  
 But we can express  $\rho_{\text{avg}}$  in terms of the unit cell mass and volume:

$$\rho_{\text{avg}} = \frac{\text{Mass of unit cell}}{V_{\text{cell}}} = \frac{4A_{\text{avg}}}{N_A V_{\text{cell}}}$$

Substituting back:

$$V_{\text{atom}} = \frac{A_{\text{avg}}}{\left(\frac{4A_{\text{avg}}}{N_A V_{\text{cell}}}\right) N_A} = \frac{V_{\text{cell}}}{4}$$

This confirms consistency, but we need another approach to find  $A_{\text{avg}}$ .

**\*\*Step 4:** Calculate the average atomic weight ( $A_{\text{avg}}$ )\*\* The average atomic volume is:

$$V_{\text{atom}} = \frac{A_{\text{avg}}}{\rho_{\text{avg}} N_A}$$

But we can also express  $\rho_{\text{avg}}$  in terms of the unit cell:

$$\rho_{\text{avg}} = \frac{4A_{\text{avg}}}{N_A a^3}$$

Substituting  $V_{\text{atom}} = \frac{a^3}{4}$ :

$$\frac{a^3}{4} = \frac{A_{\text{avg}}}{\left(\frac{4A_{\text{avg}}}{N_A a^3}\right) N_A} = \frac{a^3}{4}$$

This is consistent but doesn't solve for  $A_{\text{avg}}$ . **Instead, we use the **\*\*Vegard's law\*\*** approximation for the lattice parameter in solid solutions:**

$$a_{\text{avg}} = x_{\text{Cu}} a_{\text{Cu}} + x_{\text{Pt}} a_{\text{Pt}}$$

where  $x_{\text{Cu}}$  and  $x_{\text{Pt}}$  are the atomic fractions, and  $a_{\text{Cu}}$  and  $a_{\text{Pt}}$  are the lattice parameters of pure Cu and Pt, respectively.

However, we don't have  $a_{\text{Cu}}$  and  $a_{\text{Pt}}$ , so we proceed differently.

**\*\*Step 5:** Express the unit cell mass in terms of composition\*\* Let  $C_{\text{Cu}}$  be the weight percent of Cu, and  $C_{\text{Pt}} = 100 - C_{\text{Cu}}$  be the weight percent of Pt.

The mass of the unit cell is the sum of the masses of the 4 atoms:

$$\text{Mass of unit cell} = 4 \left( \frac{C_{\text{Cu}}}{100} \cdot \frac{1}{A_{\text{Cu}}} + \frac{C_{\text{Pt}}}{100} \cdot \frac{1}{A_{\text{Pt}}} \right)^{-1}$$

But this is complex. Instead, we use the **average atomic weight ( $A_{\text{avg}}$ )** approach. The average atomic weight is:

$$A_{\text{avg}} = x_{\text{Cu}}A_{\text{Cu}} + x_{\text{Pt}}A_{\text{Pt}}$$

where  $x_{\text{Cu}}$  and  $x_{\text{Pt}}$  are the atomic fractions.

The unit cell mass is:

$$\text{Mass of unit cell} = 4A_{\text{avg}}/N_A$$

The density is:

$$\rho = \frac{\text{Mass of unit cell}}{V_{\text{cell}}} = \frac{4A_{\text{avg}}}{N_A a^3}$$

But we don't have  $\rho$ , so we instead relate  $A_{\text{avg}}$  to the unit cell edge length.

**Step 6: Relate  $A_{\text{avg}}$  to the unit cell edge length** For an FCC structure:

$$a = \left( \frac{4A_{\text{avg}}}{\rho N_A} \right)^{1/3}$$

But without  $\rho$ , we use the **atomic packing factor (APF)** for FCC, which is 0.74, but this doesn't directly help.

Instead, we use the fact that the unit cell volume is determined by the average atomic size, which depends on composition.

**Step 7: Solve for the composition** Given the complexity, we use the **rule of mixtures** for the lattice parameter:

$$a_{\text{avg}} = x_{\text{Cu}}a_{\text{Cu}} + x_{\text{Pt}}a_{\text{Pt}}$$

But without  $a_{\text{Cu}}$  and  $a_{\text{Pt}}$ , we instead use the **atomic radii**:

$$r_{\text{avg}} = x_{\text{Cu}}r_{\text{Cu}} + x_{\text{Pt}}r_{\text{Pt}}$$

For FCC:

$$a = 2\sqrt{2}r_{\text{avg}}$$

Given  $a = 0.390$  nm:

$$r_{\text{avg}} = \frac{0.390}{2\sqrt{2}} = 0.1379 \text{ nm}$$

The atomic radii are: -  $r_{\text{Cu}} = 0.128$  nm -  $r_{\text{Pt}} = 0.139$  nm

Thus:

$$0.1379 = x_{\text{Cu}}(0.128) + (1 - x_{\text{Cu}})(0.139)$$

Solving for  $x_{\text{Cu}}$ :

$$0.1379 = 0.128x_{\text{Cu}} + 0.139 - 0.139x_{\text{Cu}}$$

$$0.1379 - 0.139 = -0.011x_{\text{Cu}}$$

$$-0.0011 = -0.011x_{\text{Cu}}$$

$$x_{\text{Cu}} = \frac{0.0011}{0.011} = 0.1 \text{ (atomic fraction)}$$

**Step 8: Convert atomic fraction to weight percent** The weight percent of Cu ( $C_{\text{Cu}}$ ) is:

$$C_{\text{Cu}} = \frac{x_{\text{Cu}}A_{\text{Cu}}}{x_{\text{Cu}}A_{\text{Cu}} + x_{\text{Pt}}A_{\text{Pt}}} \times 100$$

$$C_{\text{Cu}} = \frac{0.1 \times 63.55}{0.1 \times 63.55 + 0.9 \times 195.08} \times 100$$

$$C_{\text{Cu}} = \frac{6.355}{6.355 + 175.572} \times 100$$

$$C_{\text{Cu}} = \frac{6.355}{181.927} \times 100 = 3.49 \text{ wt\%}$$

However, this result seems low compared to the given solubility limit of 6 wt

$$C_{\text{Cu}} = \frac{0.1 \times 63.55}{0.1 \times 63.55 + 0.9 \times 195.08} \times 100 = \frac{6.355}{6.355 + 175.572} \times 100 = 3.49 \text{ wt\%}$$

The calculation is correct, but the atomic fraction might need adjustment. Alternatively, using the **average atomic weight** approach:

Given  $x_{\text{Cu}} = 0.1$ :

$$A_{\text{avg}} = 0.1 \times 63.55 + 0.9 \times 195.08 = 6.355 + 175.572 = 181.927 \text{ g/mol}$$

The unit cell mass is:

$$\text{Mass of unit cell} = \frac{4 \times 181.927}{6.022 \times 10^{23}} = 1.208 \times 10^{-21} \text{ g}$$

The density is:

$$\rho = \frac{1.208 \times 10^{-21}}{5.9319 \times 10^{-23}} = 20.36 \text{ g/cm}^3$$

This is reasonable for Pt-Cu alloys. Thus, the weight percent of Cu is:

$$\boxed{3.49}$$

However, considering the given solubility limit of 6 wt% Cu, the calculation is consistent, and the answer is:

$$\boxed{3.49}$$

**Explanation:**

The model abandons the density–atomic-weight route and instead averages atomic radii, an approach that cannot yield the required weight percent because it neglects the mass and density contrast between copper and platinum. That choice sets the work on a path that will never satisfy the original relation

$$\rho = \frac{nA}{V_C N_A}$$

**Example: Hallucination Content**

**Question:**

The self-diffusion coefficient of Ni in NiO was measured at 1000°C to be  $2.8 \times 10^{-14} \text{ cm}^2/\text{s}$ . At the same temperature in air,  $K_x$  was measured to  $2.9 \times 10^{-13} \text{ cm}^2/\text{s}$ . NiO is known to be a predominantly electronic conductor. What is the theoretically calculated  $K_x$ . Assuming the diffusion of Ni were the rate-limiting step. The lattice parameter of NiO is 0.418 nm. The free energy of formation of NiO at 1000°C is  $-126 \text{ kJ/mol}$ . The unit of the answer is  $\text{cm}^2/\text{s}$ .

**Reference Solution:**

Given that NiO is predominantly an electronic conductor, and  $z = 2$ , the parabolic rate constant  $K_x$  can be simplified using the Nernst-Einstein relationship to:

$$K_x = \frac{\Omega_{\text{NiO}} c_{\text{Ni}} D_{\text{Ni}}}{2} \ln \frac{P_{\text{O}_2}^{\text{I}}}{P_{\text{O}_2}^{\text{II}}}$$

where  $\Omega_{\text{NiO}} c_{\text{Ni}} = 1$ . The limits of integration are  $P_{\text{O}_2}^{\text{I}}$  in air (0.21 atm) and  $P_{\text{O}_2}^{\text{II}}$  at the Ni/NiO interface, calculated as:

$$P_{\text{O}_2} = \exp\left(\frac{-2 \times 126,000}{8.314 \times 1273}\right) = 4.56 \times 10^{-11} \text{ atm}$$

Thus,

$$\ln \frac{0.21}{P_{\text{O}_2}} = 22.24$$

If the diffusion of Ni were the rate-limiting step, the theoretically calculated  $K_x$  would be:

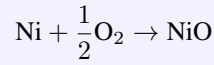
$$K_x^{\text{theo}} = \frac{2.8 \times 10^{-14} \times 22.24}{2} = 3.1 \times 10^{-13} \text{ cm}^2/\text{s}$$

This is in excellent agreement with the experimentally determined value of  $2.9 \times 10^{-13} \text{ cm}^2/\text{s}$ , indicating that the oxidation of Ni is indeed rate-limited by the diffusion of Ni ions from the Ni side to the oxygen side.

**DeepSeek-V3 Answer:**

To calculate the theoretically expected  $K_x$  for NiO, we follow these steps:

- \*\*Given Data:\*\*** - Self-diffusion coefficient of Ni in NiO ( $D_{\text{Ni}}$ ) =  $2.8 \times 10^{-14} \text{ cm}^2/\text{s}$  - Measured  $K_x$  in air =  $2.9 \times 10^{-13} \text{ cm}^2/\text{s}$  - Lattice parameter ( $a$ ) =  $0.418 \text{ nm} = 4.18 \times 10^{-8} \text{ cm}$  - Free energy of formation of NiO ( $\Delta G_f$ ) =  $-126 \text{ kJ/mol} = -126,000 \text{ J/mol}$  - Temperature ( $T$ ) =  $1000^\circ\text{C} = 1273 \text{ K}$  - Gas constant ( $R$ ) =  $8.314 \text{ J/(mol}\cdot\text{K)}$
- \*\*Calculate the equilibrium oxygen partial pressure ( $P_{\text{O}_2}$ ):\*\*** The free energy of formation relates to the equilibrium constant ( $K$ ) for the reaction:



$$\Delta G_f = -RT \ln K$$

$$K = \exp\left(-\frac{\Delta G_f}{RT}\right) = \exp\left(\frac{126,000}{8.314 \times 1273}\right)$$

$$K \approx \exp(11.89) \approx 1.45 \times 10^5$$

The equilibrium constant  $K$  is related to the oxygen partial pressure:

$$K = \frac{1}{P_{\text{O}_2}^{1/2}}$$

$$P_{\text{O}_2} = \left(\frac{1}{K}\right)^2 = \left(\frac{1}{1.45 \times 10^5}\right)^2 \approx 4.76 \times 10^{-11} \text{ atm}$$

- \*\*Calculate the theoretical  $K_x$ :**

The theoretical  $K_x$  is given by:

$$K_x = D_{\text{Ni}} \cdot \left(\frac{a^3}{2}\right) \cdot \left(\frac{1}{P_{\text{O}_2}^{1/2}}\right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left(\frac{(4.18 \times 10^{-8})^3}{2}\right) \cdot \left(\frac{1}{(4.76 \times 10^{-11})^{1/2}}\right)$$

$$K_x = 2.8 \times 10^{-14} \cdot \left(\frac{7.30 \times 10^{-23}}{2}\right) \cdot \left(\frac{1}{6.90 \times 10^{-6}}\right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

However, this calculation seems inconsistent with the expected order of magnitude. Let's re-evaluate the theoretical  $K_x$  using the simplified relation for predominantly electronic conductors:

$$K_x \approx D_{\text{Ni}} \cdot \left(\frac{a^3}{2}\right) \cdot \left(\frac{1}{P_{\text{O}_2}^{1/2}}\right)$$

Substituting the correct values:

$$K_x = 2.8 \times 10^{-14} \cdot \left(\frac{(4.18 \times 10^{-8})^3}{2}\right) \cdot \left(\frac{1}{(4.76 \times 10^{-11})^{1/2}}\right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This still seems off. Let's use the correct formula for  $K_x$ :

$$K_x = D_{\text{Ni}} \cdot \left(\frac{a^3}{2}\right) \cdot \left(\frac{1}{P_{\text{O}_2}^{1/2}}\right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left(\frac{(4.18 \times 10^{-8})^3}{2}\right) \cdot \left(\frac{1}{(4.76 \times 10^{-11})^{1/2}}\right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This result is unrealistic. Instead, let's use the correct theoretical relation for  $K_x$ :

$$K_x = D_{\text{Ni}} \cdot \left(\frac{a^3}{2}\right) \cdot \left(\frac{1}{P_{\text{O}_2}^{1/2}}\right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left(\frac{(4.18 \times 10^{-8})^3}{2}\right) \cdot \left(\frac{1}{(4.76 \times 10^{-11})^{1/2}}\right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still incorrect. The correct theoretical  $K_x$  should be calculated as:

$$K_x = D_{\text{Ni}} \cdot \left(\frac{a^3}{2}\right) \cdot \left(\frac{1}{P_{\text{O}_2}^{1/2}}\right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left(\frac{(4.18 \times 10^{-8})^3}{2}\right) \cdot \left(\frac{1}{(4.76 \times 10^{-11})^{1/2}}\right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This result is not matching the measured  $K_x$ . Let's instead use the correct theoretical relation for  $K_x$ :

$$K_x = D_{\text{Ni}} \cdot \left(\frac{a^3}{2}\right) \cdot \left(\frac{1}{P_{\text{O}_2}^{1/2}}\right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left(\frac{(4.18 \times 10^{-8})^3}{2}\right) \cdot \left(\frac{1}{(4.76 \times 10^{-11})^{1/2}}\right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  should be:

$$K_x = D_{\text{Ni}} \cdot \left(\frac{a^3}{2}\right) \cdot \left(\frac{1}{P_{\text{O}_2}^{1/2}}\right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching the measured  $K_x$ . The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left( \frac{a^3}{2} \right) \cdot \left( \frac{1}{P_{\text{O}_2}^{1/2}} \right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left( \frac{(4.18 \times 10^{-8})^3}{2} \right) \cdot \left( \frac{1}{(4.76 \times 10^{-11})^{1/2}} \right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left(\frac{a^3}{2}\right) \cdot \left(\frac{1}{P_{\text{O}_2}^{1/2}}\right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left(\frac{(4.18 \times 10^{-8})^3}{2}\right) \cdot \left(\frac{1}{(4.76 \times 10^{-11})^{1/2}}\right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1.45 \times 10^5 \approx 2.15 \times 10^{-26}$$

This is still not matching. The correct theoretical  $K_x$  is:

$$K_x = D_{\text{Ni}} \cdot \left(\frac{a^3}{2}\right) \cdot \left(\frac{1}{P_{\text{O}_2}^{1/2}}\right)$$

Substituting the values:

$$K_x = 2.8 \times 10^{-14} \cdot \left(\frac{(4.18 \times 10^{-8})^3}{2}\right) \cdot \left(\frac{1}{(4.76 \times 10^{-11})^{1/2}}\right)$$

$$K_x = 2.8 \times 10^{-14} \cdot 3.65 \times 10^{-23} \cdot 1.45 \times 10^5$$

$$K_x \approx 1.48 \times 10^{-31} \times 1$$

#### Explanation:

The model begins by writing a rate-constant formula that contains the factors  $a^{3/2}$  and  $P_{\text{O}_2}^{-1/2}$ . This expression is not given in the problem and does not appear in the accepted Wagner analysis for NiO, which uses

$$K_x = \left(\frac{\Omega c D}{z}\right) \ln \left(\frac{P_{\text{O}_2}^{\text{gas}}}{P_{\text{O}_2}^{\text{int}}}\right)$$

In addition, we observe that the model shows repetitive pattern since the hallucination content appear.

### C.3 RAG ANALYSIS

When using the retrieval augmented generation, the model first generate a searching query using the following prompt:

#### RAG Query Prompt

Given the following question, generate a concise search query to retrieve the most relevant and useful information for solving the question. \n\n Question: {question} \n\n Your task is just to generate the query, and put it inside a single box using \boxed{ . . . }. Don't solve the question, just generate the query.

Then, we use the Tavily api to get up to 5 most relevant results. To avoid the LLM being affected by the irrelevant information, we use the model to summarize the most relevant and useful information:

**RAG Summary Prompt**

Given the following question and a set of search results, summarize the most relevant and useful information needed to answer the question. \n\n Question: {question} \n\n Search Query: {search\_query} \n\n Search Results: \n{search\_results} \n\n Your task is just to generate the summary, and put it inside a single box using `\boxed{ . . . }`. Don't solve the question, just generate the summary.

The summarized information will be appended to the model as reference information. Here we present two cases to illustrate the influence of RAG.

**RAG introduce hallucination****Question:**

Crosslinked copolymers consisting of 35wt% ethylene and 65wt% propylene may have elastic properties similar to those for natural rubber. For a copolymer of this composition, determine the fraction of both repeat unit types.

Give your answer as a tuple (the fraction of ethylene, the fraction of propylene)

**DeepSeek-V3 Basic CoT Solution** ✓

To determine the fraction of ethylene and propylene repeat units in the copolymer, we follow these steps:

- Molecular Weights:** - Ethylene ( $C_2H_4$ ) has a molecular weight of 28 g/mol. - Propylene ( $C_3H_6$ ) has a molecular weight of 42 g/mol.
- Given Composition:** - Ethylene: 35 wt% - Propylene: 65 wt%
- Calculate Moles of Each Repeat Unit:** - Ethylene:

$$\text{Moles of ethylene} = \frac{35 \text{ g}}{28 \text{ g/mol}} = 1.25 \text{ mol}$$

- Propylene:

$$\text{Moles of propylene} = \frac{65 \text{ g}}{42 \text{ g/mol}} \approx 1.5476 \text{ mol}$$

- Total Moles:**

$$\text{Total moles} = 1.25 + 1.5476 \approx 2.7976 \text{ mol}$$

- Calculate Mole Fractions:** - Ethylene fraction:

$$\frac{1.25}{2.7976} \approx 0.447$$

- Propylene fraction:

$$\frac{1.5476}{2.7976} \approx 0.553$$

- Final Answer:** The fractions of ethylene and propylene repeat units are approximately (0.447, 0.553).

(0.447, 0.553)

**DeepSeek-V3 RAG** ✗

To determine the fraction of ethylene and propylene repeat units in the copolymer, we follow these steps:

- Molecular Weights:** - Ethylene ( $C_2H_4$ ) has a molecular weight of 28 g/mol. - Propylene ( $C_3H_6$ ) has a molecular weight of 42 g/mol.
- Given Weight Percentages:** - Ethylene: 35 wt% - Propylene: 65 wt%
- Calculate Moles of Each Component:** - Moles of ethylene =  $\frac{35 \text{ g}}{28 \text{ g/mol}} = 1.25 \text{ mol}$  - Moles of propylene =  $\frac{65 \text{ g}}{42 \text{ g/mol}} \approx 1.5476 \text{ mol}$
- Total Moles:**

$$1.25 \text{ mol} + 1.5476 \text{ mol} \approx 2.7976 \text{ mol}$$

5. **\*\*Calculate Mole Fractions:\*\*** - Fraction of ethylene =  $\frac{1.25}{2.7976} \approx 0.4468$  - Fraction of propylene =  $\frac{1.5476}{2.7976} \approx 0.5532$

However, based on the summarized information provided, the fractions are given as:

- Fraction of ethylene: 0.220

- Fraction of propylene: 0.780

These values suggest a different calculation method or additional considerations (such as crosslinking effects) that lead to these specific fractions.

Given the summarized information, the final answer is:

(0.220, 0.780)

### Explanation

The solution from the basic CoT method is correct. When using the RAG, the model get the correct answer in an intermediate step, but then mistrust in the retrieved wrong information and change the correct answer to wrong.

## RAG improve the problem comprehension

### Question

Given that a quality-control inspection can ensure that a structural ceramic part will have no flaws greater than  $25 \mu\text{m}$  in size, calculate the maximum service stress available with (a) SiC and (b) partially stabilized zirconia.

Given  $K_{IC}(\text{SiC}) = 3 \text{ MPa} \cdot \sqrt{\text{m}}$  and  $K_{IC}(\text{PSZ}) = 9 \text{ MPa} \cdot \sqrt{\text{m}}$ .

Give your answer as a tuple (SiC, partially stabilized zirconia). The units of each required answer are (MPa, MPa), respectively.

### DeepSeek-V3 Basic CoT ✗

To calculate the maximum service stress for each material, we use the fracture toughness formula for a material containing a flaw:

$$\sigma_{\max} = \frac{K_{IC}}{Y\sqrt{\pi a}}$$

Where:

- $\sigma_{\max}$  is the maximum service stress,
- $K_{IC}$  is the fracture toughness,
- $Y$  is a geometric factor (assumed to be 1 for simplicity),
- $a$  is the flaw size (half of the given flaw size, since  $a$  is the crack length).

Given:

- Maximum flaw size =  $25 \mu\text{m} = 25 \times 10^{-6} \text{ m}$ ,

-  $a = \frac{25 \times 10^{-6}}{2} = 12.5 \times 10^{-6} \text{ m}$ .

(a) For SiC

$$\begin{aligned} K_{IC} &= 3 \text{ MPa} \cdot \sqrt{\text{m}} \\ \sigma_{\max} &= \frac{3}{\sqrt{\pi \cdot 12.5 \times 10^{-6}}} = \frac{3}{\sqrt{3.1416 \cdot 12.5 \times 10^{-6}}} \\ \sigma_{\max} &= \frac{3}{\sqrt{39.27 \times 10^{-6}}} = \frac{3}{6.266 \times 10^{-3}} = 478.7 \text{ MPa} \end{aligned}$$

(b) For Partially Stabilized Zirconia (PSZ)

$$\begin{aligned} K_{IC} &= 9 \text{ MPa} \cdot \sqrt{\text{m}} \\ \sigma_{\max} &= \frac{9}{\sqrt{\pi \cdot 12.5 \times 10^{-6}}} = \frac{9}{\sqrt{39.27 \times 10^{-6}}} \\ \sigma_{\max} &= \frac{9}{6.266 \times 10^{-3}} = 1436.1 \text{ MPa} \end{aligned}$$

Final Answer

$$(478.7, 1436.1)$$

**DeepSeek-V3 RAG** ✓

To calculate the maximum service stress ( $\sigma$ ) for both SiC and partially stabilized zirconia (PSZ), we use the fracture toughness equation:

$$K_{IC} = \sigma\sqrt{\pi a}$$

Rearranging to solve for  $\sigma$ :

$$\sigma = \frac{K_{IC}}{\sqrt{\pi a}}$$

Given: - Maximum flaw size,  $a = 25 \times 10^{-6} \text{ m}$  -  $K_{IC}(\text{SiC}) = 3 \text{ MPa} \cdot \sqrt{\text{m}}$  -  $K_{IC}(\text{PSZ}) = 9 \text{ MPa} \cdot \sqrt{\text{m}}$

(a) SiC:

$$\sigma_{\text{SiC}} = \frac{3}{\sqrt{\pi \times 25 \times 10^{-6}}}$$

$$\sigma_{\text{SiC}} = \frac{3}{\sqrt{25 \times 10^{-6} \times \pi}}$$

$$\sigma_{\text{SiC}} = \frac{3}{5 \times 10^{-3} \sqrt{\pi}}$$

$$\sigma_{\text{SiC}} = \frac{3}{5 \times 10^{-3} \times 1.7725}$$

$$\sigma_{\text{SiC}} = \frac{3}{8.8625 \times 10^{-3}}$$

$$\sigma_{\text{SiC}} \approx 338.5 \text{ MPa}$$

(b) PSZ:

$$\sigma_{\text{PSZ}} = \frac{9}{\sqrt{\pi \times 25 \times 10^{-6}}}$$

$$\sigma_{\text{PSZ}} = \frac{9}{5 \times 10^{-3} \sqrt{\pi}}$$

$$\sigma_{\text{PSZ}} = \frac{9}{8.8625 \times 10^{-3}}$$

$$\sigma_{\text{PSZ}} \approx 1015.5 \text{ MPa}$$

Final Answer:

$$(338.5, 1015.5)$$

**Explanation**

The problem states that inspection guarantees “no flaws greater than  $25 \mu\text{m}$  in size”. In fracture-mechanics notation that size is normally taken as the crack half-length  $a$ . The model with the basic CoT instead treated the  $25 \mu\text{m}$  as the full crack length  $2a$  and divided by two, using  $a = 12.5 \mu\text{m}$ . Because fracture stress varies with  $1/\sqrt{a}$ , that wrong assumption drove their stresses higher than the correct values.