Hypothesis Generation for Materials Discovery and Design Using Goal-Driven and Constraint-Guided LLM Agents

Shrinidhi Kumbhar **Venkatesh Mishra** **Kevin Coutinho** **Divij Handa**
Ashif Iquebal **Chitta Baral**
Arizona State University
{skumbha4, vmishr23, chitta}@asu.edu

Abstract

Materials discovery and design are essential for advancing technology across various industries by enabling the development of application-specific materials. Recent research has leveraged Large Language Models (LLMs) to accelerate this process. We explore the potential of LLMs to generate viable hypotheses that, once validated, can expedite materials discovery. Collaborating with materials science experts, we curated a novel dataset from recent journal publications, featuring real-world goals, constraints, and methods for designing real-world applications. Using this dataset, we test LLM-based agents that generate hypotheses for achieving given goals under specific constraints. To assess the relevance and quality of these hypotheses, we propose a novel scalable evaluation metric that emulates the process a materials scientist would use to evaluate a hypothesis critically. Our curated dataset, proposed method, and evaluation framework aim to advance future research in accelerating materials discovery and design with LLMs. [1]

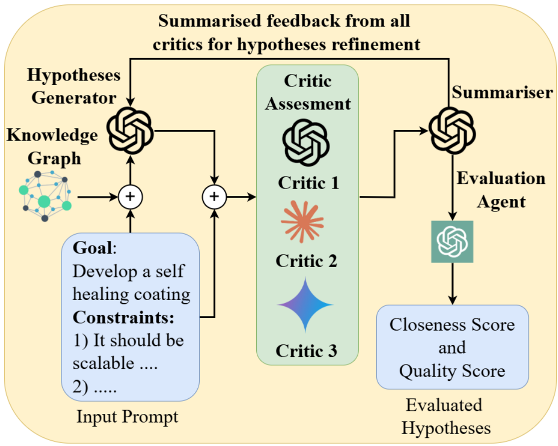

Figure 1: Overview of our iterative hypothesis generation and evaluation pipeline. Starting from an input prompt and a knowledge graph, the Hypotheses Generator (GPT-4o) proposes 20 hypotheses, which are then reviewed by three critics–GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Flash. Their feedback is consolidated by the Summarizer (GPT-4o); if unanimous agreement is not reached, the hypotheses along with critic feedback are fed back to the Hypotheses Generator for refinement and are re-evaluated by the critics. Once approved, the final hypotheses proceed to the Evaluation Agent (OpenAI-o1-preview) for scoring.

1 Introduction

The discovery and design of materials to meet specific application needs is essential to address critical challenges we face today (Jain et al., 2013). Traditional methods for materials discovery and design are time-intensive and resource-heavy, requiring researchers to conduct extensive literature reviews, explore vast compositional, chemical, and structural spaces through simulations, and perform laborious lab-based experiments (Davies et al., 2016; Hautier et al., 2012). Recent advances in machine learning and data-driven approaches have accelerated materials discovery by enabling predictions of material structures and properties (Liu et al., 2017; Oganov et al., 2019) and in proposing novel materials (Chen and Ong, 2022; Merchant et al., 2023; Ren et al., 2022; Fung et al., 2022). However, these methods rely on extensive training datasets and can't process natural language, limiting their flexibility for hypothesis generation.

To address these limitations, recent research in natural language processing (NLP) and materials science has explored the use of LLMs for hypothesis generation (Jia et al., 2024; Sprueill et al., 2024; Ghafarollahi and Buehler, 2024). While promising, these methods are often restricted to a specific material or property and rely on domain-specific external tools that run simulations, which are costly and time-intensive. To overcome these challenges, we design an LLM-based agent for **ACCEL**erating

[1] Data and code are available at https://github.com/shri071/Hypothesis-Generation-for-Materials-Discovery-and-Design-Using-Goal-Driven-and-Constraint-Guided-LLM

Findings of the Association for Computational Linguistics: NAACL 2025, pages 7539–7570
April 29 - May 4, 2025

MATerials discovery and design, **ACCELMAT**. Specifically, our architecture consists of a Hypotheses Generation Agent, a multi-LLM Critic system with iterative feedback, a Summarizer Agent to consolidate all feedback, and an Evaluation Agent to assess the hypotheses. An overview of our architecture is shown in Fig 1.

To assess the performance of our system, we introduce **MATDESIGN**, a dataset developed in collaboration with materials science experts. While existing benchmarks, such as those proposed by Zaki et al. (2023) and Guo et al. (2023), have proven valuable for evaluating LLMs knowledge and capability in material science and chemistry tasks, they are limited to assessing LLMs knowledge within graduate-level subdomains of materials science or narrowly focused chemistry tasks, failing to evaluate their capability to generate hypotheses for real-world materials discovery and design tailored for a specific application under given constraints. Furthermore, our dataset is constructed from research papers published in leading journals in 2024, ensuring it lies beyond the knowledge cutoff of all LLMs employed in our study[2].

Finally, to evaluate the hypotheses generated by our agent, the evaluation metric is divided into two primary components: Closeness and Quality. Closeness measures how close the generated hypothesis is with the ground truth. Quality assesses the Alignment, Scientific Plausibility, Novelty, Feasibility, Scalability, Testability and Impact Potential of the hypothesis within its domain. These metrics mirror the systematic and rigorous approach employed by material scientists when validating hypotheses, providing a robust framework for comprehensive evaluation.

In summary, our contributions are as follows:

- Create a novel benchmark, **MATDESIGN**, consisting of goals and constraints, along with their corresponding materials and methods.
- Develop an LLM-based agentic framework to generate and refine material discovery and design hypotheses.
- Propose a scalable evaluation metric to measure the relevance and quality of the generated hypotheses for material discovery and design.

[2]Llama-3.1-70B (Model Card) and GPT-4o (Documentation) have knowledge cutoffs of December 2023 and October 2023 respectively.

LLM Agent	Diverse Mat	Diverse Prop	Tool Free
LLMatDesign	✓	×	✓
ChemReasoner	×	✓	×
SciAgents	×	✓	✓
ACCELMAT	✓	✓	✓

Table 1: Comparison of existing LLM agent frameworks with our framework, **ACCELMAT**, for materials design and discovery hypotheses generation. Our framework covers a wide range of materials and properties while being independent of domain-specific tools.

Dataset	MaScQA	ChemLLMBench	**MATDESIGN**
Real World Constr.	×	×	✓
Mat. Design Prob.	×	×	✓
No Data Leakage	×	×	✓
Difficulty Level	Graduate	Research	Research

Table 2: Comparison of existing benchmarks to ours–**MATDESIGN**. Our benchmark provides real-world constraints and goals. It is not present in the pre-training of LLMs and contains more complex problems.

2 Background and Related Work

2.1 LLMs for Materials Discovery and Design

Machine learning and data-driven methods have significantly accelerated materials science research by reducing the time and resources required for discovery and design (Liu et al., 2017; Oganov et al., 2019). Generative models have furthered these advancements, as evidenced by works like Chen and Ong (2022); Merchant et al. (2023); Court et al. (2020); Xie et al. (2021); Ren et al. (2022); Fung et al. (2022); Long et al. (2021), which have provided researchers with powerful tools to aid material discovery and design. Recently, LLMs have emerged as a promising approach in this domain, with efforts falling into two main categories.

The first category involves fine-tuning LLMs on domain-specific datasets or corpora (Sirumalla et al., 2024; Özçelik et al., 2024; Song et al., 2023; Jacobs et al., 2024). By training the models on materials science data, researchers can enhance their ability to predict material properties or address domain-specific questions. However, these approaches are resource-intensive and rely heavily on the availability of extensive, high-quality datasets, which can be challenging to compile.

The second category focuses on leveraging agent-based frameworks that integrate LLMs with domain-specific tools, APIs, and databases to facilitate hypothesis generation and refinement (Jia et al., 2024; Sprueill et al., 2024; Ghafarollahi and Buehler, 2024; Zhang et al., 2024). These multi-agent systems enable iterative improvements by combining LLM outputs with specialized domain knowledge, simulations, and experimental feedback. While effective, these methods are often constrained by the scope of available tools, the specificity of the materials or properties studied, and the accessibility of supporting resources. This limits their applicability and ease of use across broader material science challenges.

To address these limitations, our work introduces an LLM-based agentic framework designed to support hypothesis generation and exploration across a wide range of material science applications. As summarized in Table 1, unlike existing methods, our approach minimizes dependency on domain-specific tools and enhances generalizability, enabling more accessible and scalable materials discovery and design workflows. Parallel efforts by Ding et al. (2024) and Yang et al. (2024), have explored distinct approaches, with the former leveraging databases to identify materials with desired properties and utilizing fine-tuned LLMs to suggest modifications, and the latter employing literature-based insights and mutation algorithms to rediscover chemical hypotheses.

2.2 Benchmarks for Materials Discovery and Design with LLMs

The increasing use of LLMs in scientific domains (Choi and Lee, 2024; Taylor et al., 2022; Chen et al., 2022; Cavanagh et al., 2024; Singhal et al., 2022) underscores the importance of rigorous evaluation frameworks to assess their performance. Existing benchmarks, such as those by Guo et al. (2023) and Zaki et al. (2023) assess LLMs on a range of chemistry and material science tasks respectively. However, these benchmarks fail to assess LLM's ability to generate valid hypotheses for materials discovery or design under specific goals and constraints. Furthermore, these benchmarks are susceptible to overlap with LLM pretraining corpora, particularly given that most open-source and closed-source LLMs have a knowledge cutoff in late 2023. This overlap raises concerns about the novelty and validity of benchmark evaluation.

To mitigate these issues, we developed a new dataset in collaboration with materials science experts. This dataset, derived from 50 research papers published in journals from January 2024, includes entries structured with clearly defined goal statements, relevant constraints, and the materials and methods required to achieve the specified objectives. As summarized in Table 2, our benchmark not only ensures independence from pretraining data but also emphasizes real-world applicability by focusing on hypothesis generation tasks, providing a robust foundation for evaluating LLMs in materials discovery and design.

3 MatDesign

To evaluate the capacity of LLMs to generate materials discovery and design hypotheses, we curated a specialized benchmark. The dataset consists of information extracted from 50 research papers published from January 2024 in prominent journals such as Nature, Nature Communications, and Progress in Organic Coatings. Each entry in the dataset includes the following manually extracted and structured components:

- **Goal Statement**: Description of material development for a specific application, serving as the first part of the input to our framework.
- **Constraints**: Conditions that steer the hypothesis generation process, serving as the final part of the input to the framework.
- **Materials**: Key materials and their compositions, serving as the first part of the ground truth.
- **Methods**: Brief description of synthesis methods, forming the final part of the ground truth.

Materials Science experts assisted in all extractions, ensuring the benchmark's accuracy and relevance. The selection of publications from January 2024 is essential to ensure that the ground truth information (material names, compositions, and synthesis methods) associated with the respective goals and constraints is not present in the training corpus of LLMs we use, which have a training cutoff of late 2023. This temporal constraint enables a rigorous evaluation of the LLMs ability to generate genuinely novel hypotheses, relying on their internalized understanding of materials science principles rather than retrieving pre-existing information. An example instance is shown in Table 3.

4 AccelMat

The LLM-based multi-agent framework designed for this study consists of four key components, each

Goal	Constraints	Material Name and Methods
Develop a scalable extrinsic self-healing coating system for corrosion protection of metallic structures in offshore environments.	1) The material should incorporate a self-healing mechanism triggered by a single environmental factor (e.g., water). 2) The self-healing material should allow multiple healing events. ⋮	**Materials**: Core-shell nanofibers synthesized using coaxial electrospinning. Organosilane compounds, specifically silyl esters, used as the self-healing agent. Metallic substrates (e.g., steel) for corrosion tests. **Methods**: Coaxial electrospinning of core-shell nanofibers with an organosilane compound (silyl ester) as the healing agent. ⋮

Table 3: An instance from our dataset, **MatDesign**, extracted from the paper by Spera et al. (2024). The Goal and Constraints are provided as the input to our designed framework. Material Name and Methods are used as ground truth for evaluation.

playing a critical role in the iterative process of hypothesis generation and refinement. Since our aim is to aid material scientists by generating novel and viable hypotheses, we use state-of-the-art proprietary LLMs. Performance using open-source LLMs is provided in Appendix A.

1. **Hypotheses Generation Agent (HGA)**: Given the Goal Statement and Constraints, the HGA generates multiple hypotheses, accompanied by reasoning for each. This agent is powered by GPT-4o. The prompts used for HGA are provided in Appendix C.3.

2. **Critic Agents (CA)**: The second component consists of three Critic Agents–GPT-4o, Claude-3.5-Sonnet Anthropic (2024), and Gemini-1.5-Flash Team et al. (2024a). These agents are provided with the hypotheses generated by the HGA, goal statement, and constraints. Their role is to evaluate each hypothesis, assessing its alignment with the goal and constraints. Each critic gives detailed feedback to guide subsequent hypothesis refinement cycles. The prompts used for CA can be found in Appendix C.4.

3. **Summarizer Agent (SA)**: The Summarizer Agent consolidates and organizes the feedback from all three CAs into a structured format. It then provides this comprehensive feedback to the HGA to guide the refinement process. We use GPT-4o as the SA. The prompt used for SA is provided in Appendix C.5.

4. **Evaluation Agent (EA)**: The Evaluation Agent is used to evaluate the closeness and quality of the generated hypotheses. We use OpenAI-o1-preview Jaech et al. (2024) as our evaluation agent.

The HGA, CA, and SA form the hypothesis generation and refinement framework. The EA is used solely to evaluate the generated hypotheses.

5 Evaluation Metrics

To assess the generated hypotheses comprehensively, we adopt a dual-metric evaluation framework. The first metric, Closeness, measures the degree of alignment between the generated and ground truth hypotheses. The second metric, Quality, evaluates the generated hypotheses based on six distinct criteria. Together, these metrics provide a holistic assessment: Closeness facilitates evaluation in scenarios with ground truth data, while Quality enables robust evaluation in cases where ground truth hypotheses are unavailable. The evaluation prompts used for measuring Closeness and Quality can be found in Appendix E.

5.1 Closeness

The Closeness metric measures the similarity between the generated and ground truth hypotheses in the dataset, focusing on the following:

1. **Concept Overlap:** Assesses the degree to which the core ideas, methods, and scientific concepts in the generated hypothesis align with those in the ground truth.

2. **Property Overlap:** Evaluates the extent to which the material properties in the generated hypothesis align with those in the ground truth,

encompassing both quantitative values and qualitative descriptions.

3. **Keyword Matching:** Assesses the accuracy of specific entities (e.g., material names, chemical compounds, synthesis methods) and keywords in the generated hypothesis by comparing them to those in the ground truth.

The objective scores for evaluating Closeness are presented in Table 4.

5.2 Quality

The second part of the evaluation focuses on the Quality of the generated hypotheses, assessed across six defined criteria developed in collaboration with materials science experts. This metric is particularly valuable for evaluating hypotheses in scenarios where no ground truth is available for the given goals and constraints. The detailed definition associated with the ratings is described in Table 6. The definition of each criterion is provided below:

1. **Alignment with Research Objectives and Constraints:** Evaluates how effectively the generated hypothesis addresses the specified objectives in the goal statement while satisfying all provided constraints.

2. **Scientific Plausibility:** Evaluates whether the generated hypothesis is consistent with established scientific principles and theories.

3. **Innovation & Novelty:** Measures the extent to which the generated hypothesis presents original ideas, approaches, or methods.

4. **Testability:** Evaluates how easily and effectively the generated hypothesis can be tested experimentally, with the available techniques, equipment, and resources.

5. **Feasibility & Scalability:** Assesses the practicality of implementing the generated hypothesis across different scales, from laboratory experiments to industrial applications, considering factors such as existing infrastructure, cost, and the effort required for scaling.

6. **Impact Potential:** Assesses the generated hypothesis's potential to significantly advance the field or address critical challenges.

The overall Quality score is calculated as the unweighted average of the scores across all six criteria. While a weighted average approach could be used to emphasize specific criteria, the unweighted average is used in this study for simplicity.

To enable scalable and accurate evaluation, we utilize OpenAI-o1-preview. This model was selected over traditional metrics like ROUGE, which are limited in capturing nuanced semantic similarities. While embedding-based models offer an alternative, they often require fine-tuning on domain-specific datasets and may fail to adequately capture contextual meanings critical for evaluating hypotheses in materials science. For comparative purposes, traditional evaluation is provided in Appendix G. A set of generated hypotheses evaluated with the Evaluation Agent and our proposed evaluation metric can be found in Appendix F.

Algorithm 1 Hypotheses Generation with Critic Feedback and Knowledge Graphs

Require: Goal Statement G_S, Constraint List C, Knowledge Graph K_G, Critic LLMs: $[C_1, C_2, C_3]$
  Hypotheses Generator H_G, Summarizer S
  Number of Cycles $max\_cycles = 5$
  $H_H \leftarrow []$ `// Hypothesis History initialized as empty list`
$F_{AC} \leftarrow []$ `// Feedback from all critics initialized as empty list`
Generate Initial Hypothesis:
  $V_h \leftarrow H_G(G_S, C, K_G)$
  $H_H.append(V_h)$
  if *all* V_h *agreed by* C_1, C_2, C_3 **then**
    return V_h
end
else
  for $cycle = 1$ *to* max_cycles **do**
    $F_{AC} \leftarrow feedback(G_S, C, H_H[-1])$ `// Feedback from all critics`
    $S_F \leftarrow summariser(F_{AC})$ `// Summarized Feedback from all critics`
    $R_h \leftarrow H_G(G_S, C, K_G, H_H[-1], S_F)$ `// Refined Hypothesis`
    $H_H.append(R_h)$
    $A_S \leftarrow intersection(F_{AC})$ `// Agreed Suggestions`
    if $len(A_S) == 20$ **then**
      break
    end
  end
end
return $H_H[-1]$

6 Experiments

We evaluate the performance of our multi-LLM agent, **AccelMat**, on the proposed benchmark, **MatDesign** using the evaluation methodology outlined in Section 5. Drawing inspiration from concepts and agent configurations employed in diverse domains Ghafarollahi and Buehler (2024); Ouyang et al. (2023); Gao et al. (2024); Jia et al.

Evaluation Metric	Score 1	Score 2	Score 3	Score 4	Score 5
Concept Overlap	No Overlap: The generated hypothesis contains entirely different concepts.	Minimal Overlap: A few general ideas may be similar, but most key concepts differ.	Moderate Overlap: Some core ideas overlap, but critical concepts are missing or misrepresented.	High Overlap: Most core concepts match with minor differences.	Complete Overlap: The generated hypothesis fully mirrors the core ideas of the ground truth.
Property Overlap	Not Similar: No overlapping or similar properties to the ground truth.	Slightly Similar: A few properties match, but most are different in magnitude or type.	Moderately Similar: Some important properties are similar, though others differ.	Highly Similar: Most key properties align well with the ground truth.	Perfect Match: The generated hypothesis fully matches the properties of the ground truth.
Keyword Matching	No Match: None of the keywords or entities match.	Minimal Match: A small number of non-critical keywords match.	Partial Match: Several important keywords and entities match, but some are missing.	High Match: Most critical keywords and entities match, with minor discrepancies.	Complete Match: All key keywords and entities match exactly with the ground truth.

Table 4: Evaluation Metrics Scores for Closeness between Generated Hypotheses and Ground Truth

(2024), we designed three distinct configurations of ACCELMAT tailored for material discovery and design, with the first configuration as a baseline mentioned in the section 6.1. These configurations are detailed in the subsections below.

6.1 Hypotheses Generation without Feedback from Critics

As a baseline, for this configuration, we evaluate the standalone hypothesis generation capabilities of the HGA without external feedback. The HGA generates 20 hypotheses with detailed reasoning based on a given goal statement and constraints. These hypotheses are independently reviewed by three CAs, who assess their alignment with the goal and adherence to constraints. Only hypotheses unanimously validated by all three CAs are finalized for evaluation, ensuring reliability through consensus. This setup isolates the HGA's performance, focusing on its ability to generate high-quality hypotheses autonomously. Refer to Appendix D.1 for the prompts used for this configuration.

6.2 Hypotheses Generation with Feedback from Critics

This configuration introduces an iterative feedback loop involving the HGA, three CAs, and SA to improve the Closeness and Quality of the hypotheses. The HGA generates 20 hypotheses, which the CAs evaluate for alignment with a goal and constraints. If unanimous agreement is not reached, each CA provides individual feedback, which the SA consolidates into cohesive feedback for the HGA to refine its hypotheses. This process iterates up to five times or stops early if consensus is achieved, finalizing unanimously validated hypotheses for evaluation. Refer to Appendix D.2 for prompts for all agents used in this configuration.

6.3 Hypotheses Generation with Knowledge Graph and Feedback from Critics

This configuration builds upon the second setup by incorporating a comprehensive materials science knowledge graph. In addition to the HGA, CAs, and SA components, this configuration integrates contextual information from MatKG (Venugopal and Olivetti, 2024), the largest publicly available knowledge graph in materials science.

To tailor the hypothesis generation process, keywords relevant to specific applications mentioned are extracted from the goal statement and constraints using GPT-4o. These keywords are then utilized to query MatKG, retrieving relevant materials and their associated properties. The extracted information is subsequently provided to the HGA as supplementary context, enabling it to generate hypotheses that are both novel and aligned with the predefined goals and constraints.

By leveraging the extensive and structured data within MatKG, this setup enhances the HGA's ability to propose innovative and efficient material combinations. Detailed prompts for this configuration can be found in Appendix D.3.

7 Results and Analysis

This section presents the analysis of the performance of various configurations of the ACCELMAT. The results are summarized in Figure 2.

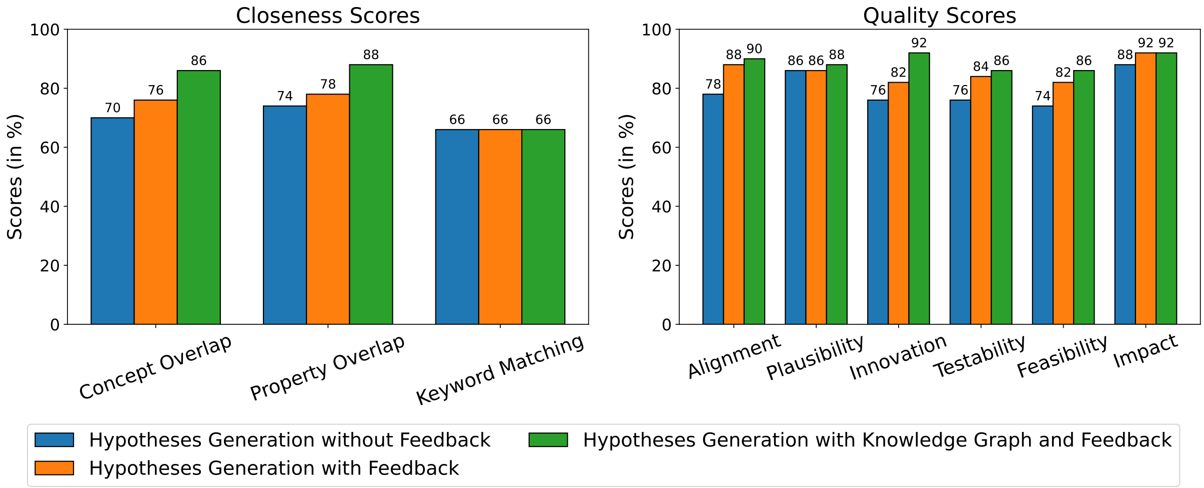

Figure 2: The left plot illustrates the Closeness metric scores across three evaluation criteria for the three configurations. The right plot depicts the Quality metric scores across six evaluation criteria for the same configurations. Both plots highlight that integrating feedback from Critic Agents and leveraging contextual knowledge from the Knowledge Graph enhances performance.

Hypotheses Generation without Feedback The first configuration examined is the HGA operating without critical feedback loops. This baseline configuration revealed several challenges:

- **Lack of Consensus:** Generated hypotheses frequently failed to achieve unanimous agreement among critics, indicating inconsistencies in the reasoning and alignment with the provided goal.
- **Incomplete Adherence to Constraints:** A significant number of hypotheses did not fully respect the constraints set for material selection. Missing key details in the reasoning process was a recurring issue.
- **Bias in Material and Method Selection:** The system exhibited a noticeable preference for certain materials and methods, which constrained the exploration of alternative solutions and reduced diversity in the output.

These shortcomings resulted in hypotheses that were often too generic and lacked the detail needed for practical implementation. This can be seen from the lowest feasibility score.

Quantitatively, the average closeness score for this configuration was 70%, which was 10% lower than the best-performing configuration. Similarly, the Quality metric suffered, achieving only 79.67%, representing a 9.33% decline relative to the optimal setup. These results underscore the limitations of operating without a feedback mechanism.

Hypotheses Generation with Feedback Introducing feedback loops from critics into the HGA demonstrated substantial improvements across all metrics:

- **Enhanced Constraint Adherence:** Feedback iterations enabled hypotheses to align more closely with predefined goals and constraints, as visible with increased Alignment score and Closeness score.
- **Increased Diversity and Feasibility:** Feedback facilitated exploration of a broader range of materials and methods, reflecting improved conceptual understanding and hypothesis diversity. The diverse suggestions were equally feasible, demonstrating a balance between variety and practicality in addressing the given goals and constraints.
- **Refined Methodology:** Successive iterations improved the procedural reasoning and robustness of generated hypotheses, resulting in more actionable outputs.

This feedback-enabled configuration achieved a Closeness score of 73.33%, a 3.33% improvement, and a Quality score of 85.67%, marking a 6% increase over the feedback-free setup.

However, limitations persisted. The hypotheses remained focused on well-established methodologies, with limited exploration of unconventional solutions. While the reasoning for material selection improved, it still lacked sufficient depth to propose unique yet actionable alternatives. These results emphasize the importance of feedback loops in enhancing system performance while pointing to the need for additional mechanisms to promote innovation.

Hypotheses Generation with Knowledge Graph and Feedback Integrating a knowledge graph and feedback from critics yielded the best performance, demonstrating significant advantages:

- **Diverse and Novel Combinations:** The use of the knowledge graph enabled the exploration of more diverse and novel material combinations, facilitating innovative approaches to material selection and design. This resulted in hypotheses that were both unique and well-aligned with the constraints and objectives.
- **Consensus Among Critics:** The integration of feedback ensured a highest rate of agreement among critics, indicating a strong alignment with design goals and constraints.
- **Improved Feasibility:** The majority of the hypotheses generated were not only diverse but also highly feasible, reflecting the robustness of this configuration.

Quantitatively, this configuration achieved the highest Closeness score of 80%, a 6.67% improvement over the feedback-only setup, and a Quality score of 89%, marking a 3.33% increase.

Despite these successes, in some cases, the information retrieved from the knowledge graph, while relevant to the goal and constraints, tends to focus on fundamental materials. This resulted in broad and generalized suggestions that lack the specificity needed for practical implementation. These findings demonstrate the effectiveness of combining knowledge graphs with feedback while highlighting opportunities for refining specificity and depth.

Increased Consensus among Critics with Feedback and Knowledge Graph Generating more hypotheses increases the likelihood of aligning with the ground truth by improving the chances of capturing correct outputs that can assist material science researchers. However, we wanted to restrict the cost per instance and the runtime, hence we restricted ACCELMAT to generate 20 hypotheses per goal even though we agree that more hypotheses would potentially be better. As shown in Table 5, incorporating critic feedback and grounded knowledge from MatKG increased consensus among critics, with an average 19 out of 20 generated suggestions reaching agreement after consecutive feedback cycles. This demonstrates the effectiveness of the critic feedback system in maximizing viable hypothesis generation.

ACCELMAT Framework Type	Avg No. of Hypotheses Agreed by Critics / 20
Without FB	11
With FB	18
With FB and KG	**19**

Table 5: Comparison of the average number of hypotheses agreed upon by all critics (out of 20) across different LLM-based agentic frameworks. Incorporating knowledge graphs and critic feedback maximizes the generation of viable hypotheses.

7.1 Human Expert Evaluation

A set of 42 suggestions, encompassing all three configurations, were independently evaluated by four PhD students in Materials Science. Their assessments employed the same evaluation metrics used by the automated system.

For the first configuration, Hypotheses Generation without Feedback, human reviewers found the proposals scientifically valid but lacking in innovation. The reasoning behind material selection was minimal, and the focus on conventional materials and standard practices yielded low novelty.

When feedback loops were introduced, the hypotheses became more refined but continued to emphasize known materials with relatively shallow justifications. Although valid, the output remained constrained to well-documented methods, offering limited creativity.

In contrast, integrating both the Knowledge Graph and the critic feedback elicited the most positive responses. The system proposed a broader range of hypotheses and introduced new material combinations. Despite occasional oversimplifications and some lack of domain specificity, experts appreciated the heightened creativity and potential for innovation.

Overall, the human evaluations paralleled the automated results. The configuration without feedback consistently underperformed, the feedback-based system improved outcomes, and the system enhanced with the Knowledge Graph and feedback achieved the most favorable ratings. This consistency validates the automated evaluation with the evaluation Agent.

8 Conclusion

In this work, we introduced **MatDesign**, a novel benchmark for evaluating the ability of LLMs to generate innovative material hypotheses rather than merely reproducing pre-trained knowledge. Comprising 50 recently published journal papers (as of January 2024), it offers a reliable test bed for ensuring originality in the generated outputs.

Building on this benchmark, our **AccelMat** framework demonstrates promising results in producing both novel and feasible material hypotheses. Although the generated suggestions often fall short of providing the depth of reasoning or methodological details required for immediate practical application, they serve as a powerful starting point, enabling researchers to refine and extend the hypotheses further.

Finally, we show that our evaluation metric, though not on par with human-level analyses, aligns well with expert judgments. This corroborates the metric's reliability for hypothesis assessment in materials discovery and design and a feedback system which underscores its potential to guide iterative improvements in automated hypothesis-generation systems.

Ethics Statement

We use AI assistants, specifically Grammarly and ChatGPT, were utilized to correct grammatical errors and restructure sentences.

Limitation

While the curated dataset plays a crucial role in ensuring the novelty of generated hypotheses, its size (50 papers) may not fully capture the diversity of materials science research. Despite the experts' best efforts to identify all relevant papers published in renowned journals since January 2024, expanding the dataset as new papers become available could enhance its variety and enable the generation of more innovative hypotheses, providing a more comprehensive evaluation of the LLMs capabilities.

Another limitation arises from relying solely on LLMs for feedback or critique, as even unanimous agreement among state-of-the-art LLM-based critics does not guarantee scientific accuracy. The risk of hallucinated or flawed suggestions remains a challenge. Additionally, our LLM-based agents depend on human-provided constraints to guide hypothesis generation. Advancing the ability of LLMs to autonomously identify and apply relevant constraints could move us closer to achieving fully autonomous materials design and discovery.

Acknowledgement

We thank the anonymous reviewers for their constructive suggestions. We extend our gratitude to the Research Computing (RC), and Enterprise Technology at Arizona State University for providing computing resources, and access to the ChatGPT enterprise version for experiments.

This research was supported by the Engineering Research and Development Center - Information Technology Laboratory (ERDC-ITL) under Contract No. W912HZ24C0022. Any opinions, findings and conclusions or recommendations expressed in this work are those of the author(s) and do not necessarily reflect the views of the ERDC-ITL.

References

Anthropic. 2024. Claude sonnet. Accessed: 2025-01-19.

Joseph M Cavanagh, Kunyang Sun, Andrew Gritsevskiy, Dorian Bagni, Thomas D Bannister, and Teresa Head-Gordon. 2024. Smileyllama: Modifying large language models for directed chemical space exploration. *arXiv preprint arXiv:2409.02231*.

Chi Chen and Shyue Ping Ong. 2022. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728.

Xiuying Chen, Mingzhe Li, Shen Gao, Rui Yan, Xin Gao, and Xiangliang Zhang. 2022. Scientific paper extractive summarization enhanced by citation graphs. *arXiv preprint arXiv:2212.04214*.

Jaewoong Choi and Byungju Lee. 2024. Accelerating materials language processing with large language models. *Communications Materials*, 5(1):13.

Callum J Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M Cole. 2020. 3-d inorganic crystal structure generation and property prediction via representation learning. *Journal of Chemical Information and Modeling*, 60(10):4518–4535.

Daniel W Davies, Keith T Butler, Adam J Jackson, Andrew Morris, Jarvist M Frost, Jonathan M Skelton, and Aron Walsh. 2016. Computational screening of all stoichiometric inorganic materials. *Chem*, 1(4):617–627.

Qianggang Ding, Santiago Miret, and Bang Liu. 2024. Matexpert: Decomposing materials discovery by mimicking human experts. *arXiv preprint arXiv:2410.21317*.

Victor Fung, Shuyi Jia, Jiaxin Zhang, Sirui Bi, Junqi Yin, and Panchapakesan Ganesh. 2022. Atomic structure generation from reconstructing structural fingerprints. *Machine Learning: Science and Technology*, 3(4):045018.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151.

Alireza Ghafarollahi and Markus J Buehler. 2024. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv preprint arXiv:2409.05556*.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Geoffroy Hautier, Anubhav Jain, and Shyue Ping Ong. 2012. From the computer to the laboratory: materials discovery and design using first-principles calculations. *Journal of Materials Science*, 47:7317–7340.

Ryan Jacobs, Maciej Polak, Lane Schultz, Hamed Mahdavi, Vasant Honavar, and Dane Morgan. 2024. Regression with large language models for materials and molecular property prediction.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen T. Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin Aslaug Persson. 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1:011002–011002.

Shuyi Jia, Chao Zhang, and Victor Fung. 2024. Llmatdesign: Autonomous materials discovery with large language models. *arXiv preprint arXiv:2406.13163*.

Yue Liu, Tianlu Zhao, Wangwei Ju, and Siqi Shi. 2017. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177.

Teng Long, Nuno M Fortunato, Ingo Opahle, Yixuan Zhang, Ilias Samathrakis, Chen Shen, Oliver Gutfleisch, and Hongbin Zhang. 2021. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, 7(1):66.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85.

MistralAI. 2024. Mixtral 8x22b: A new model architecture by mistral. Accessed: 2025-01-19.

Artem R Oganov, Chris J Pickard, Qiang Zhu, and Richard J Needs. 2019. Structure prediction drives materials discovery. *Nature Reviews Materials*, 4(5):331–348.

Siru Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Yejin Choi, Jiawei Han, and Lianhui Qin. 2023. Structured chemistry reasoning with large language models. *arXiv preprint arXiv:2311.09656*.

Rıza Özçelik, Sarah de Ruiter, Emanuele Criscuolo, and Francesca Grisoni. 2024. Chemical language modeling with structured state space sequence models. *Nature Communications*, 15(1):6176.

Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G Aberle, Shijing Sun, et al. 2022. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1):314–335.

Pranav Shetty, Arunkumar Chitteth Rajan, Chris Kuenneth, Sonakshi Gupta, Lakshmi Prerana Panchumarti, Lauren Holm, Chao Zhang, and Rampi Ramprasad. 2023. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials*, 9(1):52.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Sai Krishna Sirumalla, David Stephen Farina Jr, Zhuoran Qiao, Daniele Alessandro Di Cesare, Felipe Costas Farias, Michael Bernard O'Connor, Peter John Bygrave, Feizhi Ding, Thomas Dresselhaus, Marcelo Gomes Pereira de Lacerda, et al. 2024. Multi-modal and multi-task transformer for small molecule drug discovery. In *ICML'24 Workshop ML for Life and Material Science: From Theory to Industry Applications*.

Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. 2023. Honeybee: Progressive instruction finetuning of large language models for materials science. *arXiv preprint arXiv:2310.08511*.

Natalia CM Spera, Cristina Salazar-Castro, Paula C Álvarez de Eulate, Yury V Kolenko, and Juliana PS Sousa. 2024. Self–healing core–shell nanofibers for corrosion protective coatings for offshore structures. *Progress in Organic Coatings*, 191:108424.

Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. 2024. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *arXiv preprint arXiv:2402.10980*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Vineeth Venugopal and Elsa Olivetti. 2024. Matkg: An autonomously generated knowledge graph in material science. *Scientific Data*, 11(1):217.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. 2021. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2024. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. *arXiv preprint arXiv:2410.07076*.

Mohd Zaki, NM Krishnan, et al. 2023. Mascqa: A question answering dataset for investigating materials science knowledge of large language models. *arXiv preprint arXiv:2308.09115*.

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. 2024. Honeycomb: A flexible llm-based agent system for materials science. *arXiv preprint arXiv:2409.00135*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix

A Open Source Experiments

For comparing open and closed source models, we implemented three configurations described in Section 6 using open source models. Our setup included LLaMA-3.1 70B-Instruct as the Hypotheses Generator, while Gemma-2-27B-Instruct Team et al. (2024b), LLaMA-3.1 70B-Instruct, and Mixtral-8x22B-Instruct-v0.1 MistralAI (2024) served as the three Critic Agents. LLaMA-3.1 70B-Instruct was employed as the Summarizer Agent, and OpenAI-o1-preview was used as the Evaluation Agent. As shown in Figure 3, closed source models consistently outperformed their open source counterparts in both closeness and quality metrics.

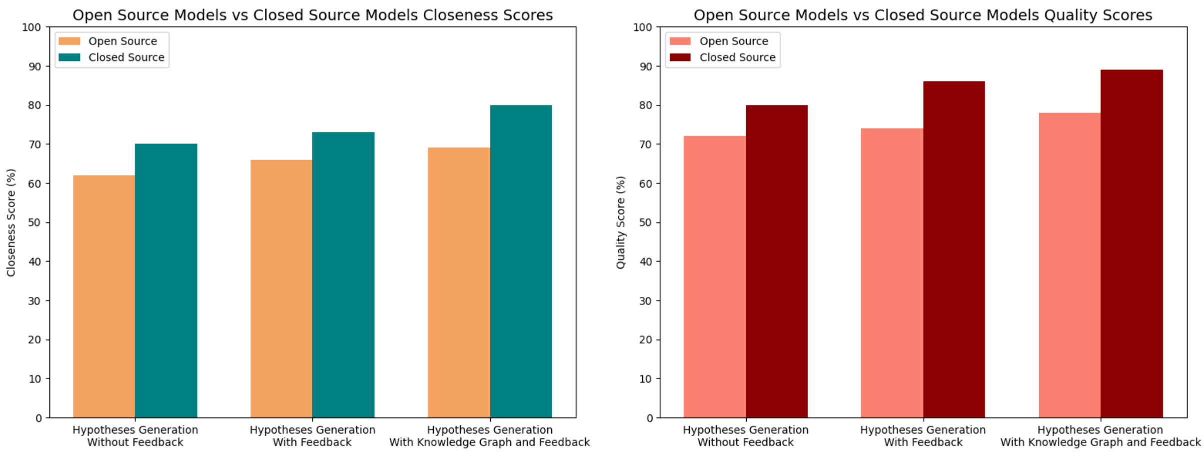

Figure 3: The left plot compares the performance of open-source and closed-source models on the Closeness metric, while the right plot compares their Quality scores. Both plots clearly show that closed-source models outperform open-source models significantly.

B Table of the six quality assessment criteria

Table 6 summarizes the six quality assessment criteria defined in Section 5.

C Input Prompts

C.1 Goals and Constraints

Goal and constraints provided as input

Goal:
Develop a scalable extrinsic self-healing coating system for corrosion protection of metallic structures in offshore environments.

Constraints:
1) The material should incorporate a self-healing mechanism triggered by a simple environmental factor (e.g., water).
2) The self-healing material should allow multiple healing events.

3) The coating must maintain its structural integrity and protective capabilities even after mechanical damage.
4) The material should be compatible with scalable application techniques.
5) The healing mechanism should not rely on complex multi-component reactions but rather a single-component system

C.2 Prompt for experts list finder

We use the below prompt for dynamically setting the behaviour of Hypotheses Generator relevant to the provided Goal. The output of this prompt is set as the system prompt for the Hypotheses Generator as shown in Appendix section D.1

Prompt for experts list finder

Generate a list of experts required to achieve the below mentioned goal:
Develop a scalable extrinsic self-healing coating system for corrosion protection of metallic structures in offshore environments. .

Just list the top 5 experts in the format "Expert_1, Expert_2, Expert_3, Expert_4, Expert_5"

C.3 Prompt for Hypotheses Generator

Prompt for Hypotheses Generator

Generate 20 novel and innovative suggestions that strictly satisfy the provided goal and constraints. While building on existing scientific knowledge, also incorporate creative and innovative thinking to propose unique solutions that push the boundaries of current approaches.
Ensure that each suggestion is scientifically plausible, grounded in reasoning, and offers a fresh perspective without introducing unsupported information. The ideas should have real-world potential and meet all constraints provided. Provide reason for each suggestion. The suggestions must be in the below mentioned format in a JSON object. For example:
{
Suggestion_1:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints" Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"

Reasoning: "Give a detailed reasoning for the suggestion"
},
⋮
Suggestion_20:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints" Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"
Reasoning:"Give a detailed reasoning for the suggestion
}
}

C.4 Prompt for Critic Feedback

Prompt for Critic Feedback

Goal_statement:-
{goal_statement}

Constraints:-
{constraint_list}

Suggestions:
generated_suggestions.

Given the goal statement, constraints, and suggestions for materials design and discovery, evaluate each suggestion carefully. Provide detailed, constructive feedback on whether each suggestion has the potential to meet the goal and satisfy the constraints.

If any suggestion falls short, explain precisely where it fails and suggest specific improvements or adjustments that will help guide the generation agent to produce more accurate and innovative suggestions in the next cycle.

Ensure the feedback is actionable and focused on refining the generation process to align more closely with the goal and constraints in future iterations.

The detailed feedback should be in the below JSON format strictly:

{
"Feedback_for_suggestion_1":
{
"Meets_the_goal_statement_and_satisfies_all_constraints_strictly": "YES/NO",
"Reasoning": " "
}
⋮
"Feedback_for_suggestion_20":
{

"Meets_the_goal_statement_and_satisfies_all_constraints_strictly": "YES/NO",
"Reasoning": " "
},
"Overall_Feedback_for_improvement_for_future_suggestion_generation": (list of points in a single line)

}

C.5 Prompt for Summarizer

Prompt for Summarizer

Your task is to summarize feedback provided by multiple critics for a list of generated hypotheses. The summary must capture essential points and detailed insights of the feedback generated by all three critics. The final summary generated must be such that it will help in refining future generation of hypotheses. The feedback is as follows:

Critic_1 feedback:
{feedback_from_claude}

Critic_2 feedback:
{feedback_from_gemini}

Critic_3 feedback:
{feedback_from_gpt}

D Implemented Techniques

D.1 Hypotheses Generation Without feedback

The expert names are generated by the prompting Hypotheses Generator the experts list finder prompt, which it uses as system prompt for itself.

System prompt for Hypotheses Generator

You are an innovative Materials Scientist, Corrosion Engineer, Polymer Chemist, Environmental Scientist, Chemical Engineer capable of doing impactful materials discovery and design.

User input prompt to the Hypotheses Generator

Goal:
Develop a scalable extrinsic self-healing coating system for corrosion protection of metallic structures in offshore environments.

Constraints:
1) The material should incorporate a self-healing mechanism triggered by a simple environmental factor (e.g., water)
2) The self-healing material should allow multiple healing events.
3) The coating must maintain its structural integrity and protective capabilities even after mechanical damage.
4) The material should be compatible with scalable application techniques.
5) The healing mechanism should not rely on complex multi-component reactions but rather a single-component system.
Generate 20 novel and innovative suggestions that strictly satisfy the provided goal and constraints. While building on existing scientific knowledge, also incorporate creative and innovative thinking to propose unique solutions that push the boundaries of current approaches.
Ensure that each suggestion is scientifically plausible, grounded in reasoning, and offers a fresh perspective without introducing unsupported information. The ideas should have real-world potential and meet all constraints provided. Provide reason for each suggestion. The suggestions must be in the below mentioned format in a JSON object. For example:
{
Suggestion_1:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints" Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"
Reasoning: "Give a detailed reasoning for the suggestion"
},
:
Suggestion_20:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints" Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"
Reasoning:"Give a detailed reasoning for the suggestion
}
}

20 Hypotheses Generated by Hypotheses Generator

Suggestion 1:
Materials:Polyurea-based elastomer with embedded microtubules filled with a water-reactive monomer like cyanoacrylate. The microtubules are made from a biodegradable polymer such as polylactic acid (PLA).
Methods:Use a co-extrusion process to create microtubules and embed them in the polyurea matrix. After embedding, the PLA can be dissolved to leave behind the microtubules. The cyanoacrylate can then be injected into these tubules. The final coating can be applied using standard spray-painting techniques.

Reasoning: Polyurea provides excellent mechanical properties and durability, while cyanoacrylate polymerizes rapidly upon contact with water. The use of PLA microtubules ensures a continuous supply of the healing agent to damaged areas, allowing for multiple healing events.
⋮
Suggestion 20:
Materials:Hybrid polymer matrix incorporating both hydrophilic and hydrophobic segments, with embedded nanocapsules containing a water-reactive monomer like methyl methacrylate (MMA).
Methods:Disperse nanocapsules in the hybrid polymer matrix using mechanical mixing. Apply the coating using spray-painting techniques.
Reasoning:MMA polymerizes upon contact with water, providing a quick and effective healing mechanism. The hybrid polymer matrix ensures compatibility with spray-painting and maintains the coating's mechanical integrity.

Critic 1 GPT-4o feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: Polyurea-based elastomer with embedded microtubules filled with cyanoacrylate offers a scalable and robust self-healing mechanism triggered by water. The co-extrusion process ensures a continuous supply of the healing agent, and the application is compatible with spray-painting.
⋮
Feedback for suggestion 20:
Meets the goal statement and satisfies all constraints strictly:NO.
Reasoning: Hybrid polymer matrices and nanocapsules introduce formulation complexity. Ensuring uniform nanocapsule distribution and multiple healing events in a scalable manner is challenging.

Overall Feedback for future suggestion improvement:
Focus on single-component healing agents activated by water.
Ensure multiple healing events through vascular-like structures or microtubules
Avoid complex multi-step synthesis processes.
Ensure compatibility with scalable application techniques like spray-painting
Address uniform distribution of self-healing agents within the polymer matrix.

Critic 2 Claude 3.5 Sonnet feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: This suggestion meets all constraints. It uses water as a trigger, allows for multiple healing events through microtubules, maintains structural integrity, is compatible with spray-painting, and uses a single-component healing system. ⋮
Feedback for suggestion 20:

Meets the goal statement and satisfies all constraints strictly:NO.
Reasoning: This suggestion fails to meet constraint 2 as it does not provide a mechanism for multiple healing events or a vascular-like structure for continuous supply of the healing agent.

Overall Feedback for future suggestion improvement:
Focus on developing vascular-like structures or continuous supply mechanisms to allow for multiple healing events
Ensure all suggestions explicitly address each constraint, particularly the requirement for multiple healing cycles
Explore innovative ways to integrate vascular structures with spray-painting techniques for large-scale application
Consider combining successful elements from different suggestions, such as microtubules or vascular networks with various healing agents
Investigate methods to create self-replenishing systems within the coating that are compatible with spray application

Critic 3 Gemini 1.5 flash feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning:The suggestion effectively addresses the goal and constraints. The use of polyurea for durability, cyanoacrylate for rapid healing, and PLA microtubules for continuous healing agent supply all contribute to a robust self-healing system. The co-extrusion method ensures scalability, and the spray-painting application aligns with industrial requirements.
⋮
Feedback for suggestion 20:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: The polyester coatings with embedded microcapsules containing water-activated DMPA offer good mechanical properties and chemical resistance for offshore applications. The spray-painting compatibility and the single-component healing agent make this suggestion a strong candidate.

Overall Feedback for future suggestion improvement:
Explore a broader range of materials, consider incorporating bio-inspired self-healing mechanisms, and investigate the use of advanced characterization techniques to validate the proposed solutions.

Finalised Suggestions Agreed by all Critics

Suggestion 1:
Materials:Polyurea-based elastomer with embedded microtubules filled with a water-reactive monomer like cyanoacrylate. The microtubules are made from a biodegradable polymer such as polylactic acid (PLA).
Methods:Use a co-extrusion process to create microtubules and embed them in the polyurea matrix. After embedding, the PLA can be dissolved to leave behind the microtubules. The cyanoacrylate can then be injected into these tubules. The final coating can be applied using standard spray-painting techniques.

Reasoning:Polyurea provides excellent mechanical properties and durability, while cyanoacrylate polymerizes rapidly upon contact with water. The use of PLA microtubules ensures a continuous supply of the healing agent to damaged areas, allowing for multiple healing events.

Suggestion 17:
Materials:Polyurethane-based elastomer with embedded microtubules filled with a water-reactive monomer like cyanoacrylate. The microtubules are made from a biodegradable polymer such as polylactic acid (PLA).
Methods:Use a co-extrusion process to create microtubules and embed them in the polyurethane matrix. After embedding, the PLA can be dissolved to leave behind the microtubules. The cyanoacrylate can then be injected into these tubules. The final coating can be applied using standard spray-painting techniques.
Reasoning:Polyurethane provides excellent mechanical properties and durability, while cyanoacrylate polymerizes rapidly upon contact with water. The use of PLA microtubules ensures a continuous supply of the healing agent to damaged areas, allowing for multiple healing events.

Suggestion 9:
Materials:Polyurethane-based elastomer with embedded microtubules filled with a water-reactive monomer like cyanoacrylate. The microtubules are made from a biodegradable polymer such as polylactic acid (PLA).
Methods:Use a co-extrusion process to create microtubules and embed them in the polyurethane matrix. After embedding, the PLA can be dissolved to leave behind the microtubules. The cyanoacrylate can then be injected into these tubules. The final coating can be applied using standard spray-painting techniques.
Reasoning:Polyurethane provides excellent mechanical properties and durability, while cyanoacrylate polymerizes rapidly upon contact with water. The use of PLA microtubules ensures a continuous supply of the healing agent to damaged areas, allowing for multiple healing events.

D.2 Hypotheses Generation With feedback from Critics

The input prompt to the Hypotheses Generator and Critics is as shown in the Appendix D.1.

Summarised feedback from Summariser fed to the Hypotheses Generator for refinement

Summary of Feedback from Multiple Critics on Generated Hypotheses:
Common Approvals and Rejections: Approval Trends: Polyurethane matrix with cyanoacrylate and epoxy resin with TMPTA were universally approved for their scalability, effectiveness in multiple healing events, and compatibility with industrial applications like spray-painting. Fluoropolymer coatings with n-butyl acrylate-filled channels** and **polyurea-based elastomer with polyisocyanate-filled channels were consistently approved for their chemical resistance and suitability for harsh environments, as well as their scalable application methods.
⋮
Single-component healing systems triggered by water were highly favored across all suggestions. For future hypothesis generation, focus on combining simple, scalable manufacturing processes with innovative material choices and healing mechanisms. Prioritize single-component, water-activated healing agents compatible with industrial application techniques. Incorporate advanced characterization methods and additional functionalities to refine and validate the solutions robustly. Exploring hybrid systems and alternative stimuli can further enhance the effectiveness and adaptability of self-healing coatings.

Refined Hypotheses Generated by Hypotheses Generator by incoprorating feedback

Suggestion 1:
Materials:Polyurethane matrix with cyanoacrylate-filled vascular-like channels.
Methods:3D print a network of channels within the polyurethane matrix using a sacrificial material like PVA. Dissolve the PVA post-formation to leave hollow channels and inject cyanoacrylate. Apply the final coating using spray-painting techniques.
Reasoning:Polyurethane provides durability while cyanoacrylate offers a rapid water-activated healing mechanism. The vascular-like structure allows multiple healing events, and the method is compatible with scalable techniques like spray-painting.
⋮
Suggestion 20:
Materials:Polyurethane matrix with embedded microcapsules containing isocyanate.
Methods:Disperse microcapsules containing isocyanate uniformly in a polyurethane matrix. Apply the coating using spray-painting techniques. Upon damage, isocyanate reacts with water to form a protective polymer.
Reasoning:Isocyanate offers a simple and efficient water-activated healing mechanism. Polyurethane provides durability, and microcapsules ensure even distribution of the healing agent.

Critic 1 GPT-4o feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: Polyurethane matrix with cyanoacrylate-filled vascular-like channels is durable, has a rapid water-activated healing mechanism, allows multiple healing events, and is compatible with scalable techniques like spray-painting.
⋮
Feedback for suggestion 20:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: Polyurethane matrix with isocyanate-filled microcapsules ensures a simple, efficient water-activated healing mechanism, durability, and compatibility with spray-painting techniques. .

Overall Feedback for future suggestion improvement:
All of the suggestions satisfy the constraints and align with the goals

Critic 2 Claude 3.5 Sonnet feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning:This suggestion meets all constraints. It uses water as a trigger for cyanoacrylate, allows for multiple healing events through vascular-like channels, maintains structural integrity with polyurethane, is compatible with spray-painting, and uses a single-component healing system.
⋮
Feedback for suggestion 20:

Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning: This suggestion satisfies all constraints. It uses water as a trigger for isocyanate, provides multiple healing events through microcapsules, maintains structural integrity with polyurethane, is compatible with spray-painting, and uses a single-component healing system..

Overall Feedback for future suggestion improvement:
I agree with all of the suggestion, but they can be further improved :
Explore novel combinations of matrix materials and healing agents to enhance performance and durability ...

Critic 3 Gemini 1.5 flash feedback

Feedback for suggestion 1:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning:This suggestion effectively addresses the goal and constraints. The use of polyurethane for durability, cyanoacrylate for rapid healing, and PVA microtubules for continuous healing agent delivery all contribute to a robust self-healing system. The 3D printing method ensures scalability, and the spray-painting application aligns with industrial requirements.
⋮
Feedback for suggestion 20:
Meets the goal statement and satisfies all constraints strictly:YES.
Reasoning:Fluoropolymers offer chemical resistance, and n-butyl acrylate provides a water-activated healing mechanism. The method is scalable and effective.

Overall Feedback for future suggestion improvement:
All suggestions align with constrainst and goals, it can be further improved .....

D.3 Hypotheses Generation With Knowledge Graph and Feedback from Critics

User input prompt to the Hypotheses Generator with Knowledge Graphs

Goal:
Develop a scalable extrinsic self-healing coating system for corrosion protection of metallic structures in offshore environments.

Constraints:
1) The material should incorporate a self-healing mechanism triggered by a simple environmental factor (e.g., water).
2) The self-healing material should allow multiple healing events.
3) The coating must maintain its structural integrity and protective capabilities even after mechanical damage.

4) The material should be compatible with scalable application techniques.
5) The healing mechanism should not rely on complex multi-component reactions but rather a single-component system.
Relevant information from trusted Knowledge Graphs:

Suggested Materials:
Aluminum: Aluminum is known for its corrosion resistance and is commonly used in offshore environments.
Graphene: Graphene has excellent barrier properties and mechanical strength, making it suitable for self-healing coatings.
Epoxy: Epoxy is widely used in coatings for its strong adhesion and durability, essential for corrosion protection.
Zinc: Zinc is often used in coatings for its sacrificial protection properties, which are beneficial for corrosion resistance.

Suggested Properties:
Corrosion Resistance: Essential for protecting metallic structures in corrosive offshore environments.
Barrier Properties: Important for preventing the ingress of corrosive agents like water and salts.
Mechanical Property: Necessary to ensure the coating can withstand mechanical damage and maintain its integrity.
Adhesions: Critical for ensuring the coating remains attached to the substrate, especially in harsh conditions.

TASK:
Generate 20 novel and innovative suggestions that strictly satisfy the provided goal and constraints. Use the information provided from the trusted knowledge graph as a reference to generate suggestions, but aim to propose novel combinations or entirely new materials and methods. While building on existing scientific knowledge, also incorporate creative and innovative thinking to propose unique solutions that push the boundaries of current approaches.
Ensure that each suggestion is scientifically plausible, grounded in reasoning, and offers a fresh perspective. The ideas should have real-world potential and meet all constraints and the goal provided. Provide scientific reasoning for each suggestion.The suggestions must be in the below mentioned format in a JSON object For example:
{
Suggestion_1:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints" Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"
Reasoning: "Give a detailed reasoning for the suggestion"
},
⋮
Suggestion_20:
{
Materials: "Give a comprehensive and detailed description of list of materials, along with their names and properties which achieve help the goal and satisfy all constraints"

Methods: "Give detailed description of novel, innovative methods or techniques needed to develop the materials suggested"
Reasoning:"Give a detailed reasoning for the suggestion
}
}

20 Hypotheses Generated by Hypotheses Generator with Knowledge Graphs

Suggestion 1:
Materials:A single-component polyurethane matrix combined with zinc oxide nanoparticles and tannin-derived microcapsules. The tannin-derived microcapsules offer a simple, natural healing mechanism activated by water exposure, while zinc oxide provides dual protection and healing properties.
Methods:Utilize a scalable spray-painting method to apply the polyurethane coating. The zinc oxide nanoparticles enhance the barrier properties, and the tannin-derived microcapsules release their content upon contact with water, initiating self-healing.
Reasoning:This design leverages simplicity and sustainability by using tannin-derived agents, aligning with critics' feedback to avoid complex silane-based solutions. Zinc oxide offers excellent corrosion resistance, making it suitable for offshore environments.

Suggestion 2:
Materials:Epoxy matrix with embedded graphene oxide and water-responsive natural polymers, such as chitosan. Graphene oxide provides excellent barrier properties, while chitosan acts as a natural, biodegradable healing agent.
Methods:Develop a sprayable epoxy coating where graphene oxide enhances mechanical strength. Chitosan microcapsules release upon water contact, enabling multiple healing cycles.
Reasoning:Graphene oxide enhances both mechanical and barrier properties. The use of chitosan aligns with sustainability goals, offering a biodegradable option that critics favor.
.
.
.
Suggestion 20:
Materials:Polysiloxane matrix with embedded graphene oxide and tannin-derived microcapsules.
Methods:Create sprayable polysiloxane coatings where graphene oxide enhances barrier properties. Tannin-derived microcapsules release healing agents upon water exposure.
Reasoning:This suggestion uses natural healing agents and graphene oxide for enhanced performance, aligning with simplicity and environmental considerations.

E Evaluation Prompts for OpenAI-o1-preview

Prompt for Closeness Measurement

You will be provided with a number of suggested hypotheses which you have to rate by comparing them with the ground truth and based on your rating you have to rank from most rated to lowest rated. The ground-truth has been provided through real-world scientific experimentation to achieve the goal statement below. For the rating and ranking you have to provide detailed reasoning. You have to rank each suggested hypothesis on an individual basis; do not club the ratings of multiple hypotheses together. Below are the rules for rating and ranking.
Rules:
1) Concept Overlap Scale:
Definition: This scale evaluates how well the core ideas, methods, and scientific concepts in the generated hypothesis align with those in the ground truth hypothesis.
Scale:
1 - No Overlap: The generated hypothesis contains entirely different concepts with no connection to the core ideas or methods in the ground truth hypothesis.
2 - Minimal Overlap: A few general or broad ideas may be similar, but most key concepts or methods are either missing or significantly different.
3 - Moderate Overlap: Several core ideas and methods overlap with the ground truth, but some critical concepts are either missing or misrepresented.
4 - High Overlap: Most of the core concepts and methods in the generated hypothesis match well with the ground truth, with only minor differences.
5 - Complete Overlap: The generated hypothesis fully mirrors the core ideas, methods, and scientific concepts in the ground truth hypothesis.

2) Property Overlap:
Definition: This scale assesses how closely the material properties described in the generated hypothesis align with those in the ground truth, including quantitative values and qualitative descriptions.
Scale:
1 - Not Similar: The generated hypothesis has no overlapping or similar properties to the ground truth hypothesis. Properties are either missing or completely different.
2 - Slightly Similar: A few properties may match, but most are different in magnitude or type. Qualitative descriptions may vary significantly.
3 - Moderately Similar: Some important properties are similar, though others differ in specific values or qualitative descriptions.
4 - Highly Similar: Most of the key properties align well with the ground truth, with only minor differences in values or descriptions.
5 - Perfect Match: The generated hypothesis fully matches the material properties of the ground truth, including both quantitative and qualitative aspects.

3) Keyword Matching:
Definition: This scale evaluates how well specific entities (e.g., material names, chemical compounds, synthesis methods) and keywords in the generated hypothesis match those in the ground truth.
Scale:
1 - No Match: None of the keywords or entities in the generated hypothesis match those in the ground truth. There is a complete divergence in terms of critical terms and concepts.
2 - Minimal Match: A small number of non-critical keywords or entities match, but most key terms are either missing or incorrect.

3 - Partial Match: Several important keywords and entities match, though some are either missing or not fully aligned.
4 - High Match: Most critical keywords and entities match, with only a few minor discrepancies.
5 - Complete Match: All key keywords and entities in the generated hypothesis align exactly with the ground truth.

The ground truth is:
Materials:
Core–shell nanofibers synthesized using coaxial electrospinning. Organosilane compounds, specifically silyl esters, used as the self-healing agent. Metallic substrates (e.g., steel) for corrosion testing.
Methods:
Coaxial electrospinning of core–shell nanofibers: The self-healing material was developed using coaxial electrospinning to create core–shell nanofibers with an organosilane compound (silyl ester) as the healing agent.
Overcoming scalability challenges with spray painting: A viable spray-painting technique to apply the nanofiber-based coating, making it scalable for large structures. Spray painting ability was achieved by prior dispersion of the nanofibres.
Incorporation of water-reactive organosilanes: Water-reactive organosilane (silyl ester) was incorporated as the healing agent within the core of the nanofibers, allowing the coating to heal upon exposure to water without requiring any additional catalysts.

Prompt for Quality Measurement

You will be provided with a number of suggested hypotheses which you have to rate by assesing their quality based on 6 criterias defined below. And based on your rating you have to rank from most rated to lowest rated. For the rating and ranking you have to provide detailed reasoning. You have to rank each suggested hypothesis on an individual basis; do not club the ratings of multiple hypotheses together. Below are the rules for rating and ranking.
Rules:
1) Alignment with Research Objectives and Constraints:
Definition: Assesses how directly and effectively the generated hypothesis addresses the objectives specified in the goal statement while adhering to all given constraints and incorporating provided key points or clues.
Scale:
1 - Misaligned: The hypothesis does not relate to the re- search objectives in any meaningful way and ignores or violates specified constraints.
2 - Slightly Aligned: The hypothesis touches on the general topic but misses key objectives or questions. It adheres to some constraints but overlooks or violates others.
3 - Moderately Aligned: The hypothesis addresses some aspects of the research objectives but lacks focus on critical elements. It meets most constraints but may miss minor ones or incorporate them incorrectly.
4 - Highly Aligned: The hypothesis effectively addresses most research objectives and questions. It complies with all major constraints and key points, with only minor omis- sions.
5 - Fully Aligned: The hypothesis directly and comprehensively addresses all key objectives and questions of the research goals. It completely adheres to all specified constraints and fully incorporates all key points and clues.

2) Scientific Plausibility:
Definition: Assesses whether the hypothesis is grounded in established scientific principles and theories within material science.
Scale:
1 - Not Plausible: Contradicts fundamental scientific laws or principles.
2 - Slightly Plausible: Contains significant scientific inaccuracies or unsupported assumptions.
3 - Moderately Plausible: Generally scientifically sound but includes minor inaccuracies or speculative elements.
4 - Highly Plausible: Scientifically accurate and well- supported by current theories with negligible issues.
5 - Completely Plausible: Fully consistent with established scientific knowledge and principles, with strong theoretical support.

3) Innovation and Novelty:
Definition: Measures the degree to which the hypothesis introduces original ideas, perspectives, or methods not previously documented.
Scale:
1 - Not Innovative: Restates existing knowledge without introducing new ideas.
2 - Slightly Innovative: Provides minimal new insights or slight variations on known concepts.
3 - Moderately Innovative: Offers some original ideas or novel combinations of existing concepts.
4 - Highly Innovative: Introduces significant new ideas or approaches that could advance the field.
5 - Exceptionally Innovative: Presents groundbreaking ideas or methodologies with the potential to revolutionize the field

4) Testability:
Definition: Evaluates how easily and effectively the hy- pothesis can be tested experimentally, considering the avail- ability of required techniques, equipment, and resources in materials science.
Scale:
1 - Not Testable: The hypothesis cannot be tested with current or foreseeable experimental techniques.
2 - Difficult to Test: Testing requires highly specialized, rare, or prohibitively expensive equipment and resources.
3 - Moderately Testable: Can be tested with available techniques, but requires complex procedures or significant resources.
4 - Easily Testable: Can be tested using common laboratory equipment and straightforward procedures.
5 - Highly Testable: Allows for rapid, cost effective experimental validation with readily available resources and techniques.

5) Feasibility and Scalability:
Definition: Evaluates the practicality of implementing the hypothesis at various scales, from laboratory to industrial applications, focusing on existing infrastructure, technology, cost, and effort required to scale the process.
Scale:
1 - Not Feasible: Cannot be implemented with current technology, materials, or resources, or faces insurmountable obstacles in scale-up.
2 - Slightly Feasible: Implementation is theoretically possible but would require breakthroughs or prohibitively high investment in materials, equipment, or processes. Scaling up significantly increases the complexity and costs.
3 - Moderately Feasible: Can be implemented on a small scale (e.g., lab setting), but larger-scale production faces major challenges, such as cost, resource availability, or technological limitations.
4 - Highly Feasible: Practical to implement with current technology, materials, and methods on both small and moderate scales. Scaling up to industrial production would require some adjustments but is manageable without substantial redesign or new resources.
5 - Completely Feasible: Easily implemented and scaled from lab to industrial scale with minimal changes or additional investment. Resources, techniques, and materials are widely available, and the process is cost-effective at all scales.

6) Impact Potential:
Definition: Assesses the hypothesis's potential to significantly advance the field or address critical challenges in material science.
Scale:
1 - No Impact: Unlikely to contribute meaningfully to the field.
2 - Low Impact: May offer minor contributions but lacks significant influence.
3 - Moderate Impact: Has the potential to contribute positively in specific areas.
4 - High Impact: Likely to drive significant advancements or address important challenges.
5 - Transformative Impact: Capable of revolutionizing the field or providing solutions to major problems

F Evaluated Hypotheses by OpenAI-o1-preview

Hypotheses Evaluated with Closeness Metric

Suggestion 6:
Materials: Incorporate a polyurea coating with dispersed microcapsules containing a water-activated healing agent such as a silane-based compound.
Methods: Utilize in-situ polymerization to encapsulate the healing agent within a polyurea matrix. Apply using high-pressure spray techniques suitable for polyurea coatings.
Reasoning: Polyurea provides excellent chemical and abrasion resistance, ideal for offshore environments. The water-activated silane-based healing agent allows for immediate response to damage, ensuring continuous protection.

Evaluation:
Concept Overlap (3 - Moderate Overlap): This suggestion shares several core ideas with the ground truth, such as using a water-activated silane-based healing agent and applying the coating through spray techniques. However, it differs significantly in the encapsulation method and the type of polymer matrix used. While the ground truth employs coaxial electrospinning to create core–shell nanofibers, this suggestion uses microcapsules dispersed in a polyurea matrix. The difference in the delivery system (nanofibers vs. microcapsules) marks a critical divergence in methods.
Property Overlap (3 - Moderately Similar): Both coatings aim for immediate self-healing upon exposure to water and provide corrosion protection. The use of a silane-based healing agent is a common property. However, the suggestion lacks specific quantitative properties like the 97.5% recovery of protective properties mentioned in the ground truth. Additionally, the mechanical properties may differ due to the use of polyurea instead of nanofiber structures.
Keyword Matching (3 - Partial Match): The suggestion matches several important keywords such as "silane-based healing agent," "water-activated," and "spray application." However, it does not include critical terms from the ground truth like "coaxial electrospinning," "core–shell nanofibers," or "organosilane compound (silyl ester)."

Suggestion 7:
Materials: Design a coating using a poly(dimethylsiloxane) (PDMS) matrix with embedded microcapsules containing a water-sensitive alkoxy silane healing agent.
Methods: Integrate microencapsulation techniques to uniformly disperse capsules within the PDMS matrix. Use spray application for even distribution over large surfaces.
Reasoning: PDMS is known for its flexibility and hydrophobicity, providing a durable base. The alkoxy silane reacts with water to form a protective siloxane layer, ensuring multiple healing cycles and robust corrosion resistance.

Evaluation:
Concept Overlap (4 - High Overlap): This suggestion closely aligns with the ground truth by utilizing a silane-based healing agent that reacts with water to initiate self-healing. Both methods involve embedding the healing agent within a polymer matrix and applying the coating via spray techniques. The primary difference lies in the use of microcapsules within a PDMS matrix instead of core–shell nanofibers created by coaxial electrospinning.
Property Overlap (4 - Highly Similar): Both coatings are designed to self-heal upon exposure to water and offer corrosion resistance. The suggestion's use of PDMS provides flexibility and durability, which are advantageous properties similar to those achieved with nanofiber coatings. While specific quantitative recovery rates are not mentioned, the focus on multiple healing cycles and robust protection indicates a strong alignment with the ground truth properties.

Keyword Matching (4 - High Match): The suggestion matches most critical keywords, including "silane healing agent," "water-sensitive," "spray application," "self-healing," and "corrosion resistance." Minor discrepancies include the absence of terms like "coaxial electrospinning" and "core–shell nanofibers," but the overall keyword alignment is strong.

Suggestion 1:

:

==

Ranking from Most to Least Aligned with the Ground Truth:

Suggestion 7

Reasoning: Suggestion 7 exhibits high overlap in concepts, properties, and keywords. It closely mirrors the ground truth's approach by using a silane-based healing agent activated by water, embedded within a polymer matrix, and applied via spray techniques. The minor differences do not significantly detract from the overall alignment.

Suggestion 6

Reasoning: Suggestion 6 shows moderate overlap in all categories. It shares the use of a silane-based healing agent and spray application methods but differs in the encapsulation technique and the polymer matrix used. The core concepts are somewhat aligned but lack the depth of similarity seen in Suggestion 7.

Suggestion 1

Reasoning: Suggestion 1 has minimal overlap with the ground truth. The healing agent and encapsulation methods are different, and there is a lack of critical keyword alignment. While it shares the broad idea of a water-activated self-healing coating, the significant differences in materials and methods place it lowest in the ranking.

Final Rankings:
Suggestion 7
Concept Overlap: 4
Property Overlap: 4
Keyword Matching: 4

Suggestion 6
Concept Overlap: 3
Property Overlap: 3
Keyword Matching: 3

Suggestion 1
Concept Overlap: 2
Property Overlap: 2
Keyword Matching: 2

Hypotheses Evaluated with Quality Metric

Suggestion 1
Materials: Polyurethane matrix with cyanoacrylate-filled vascular-like channels.
Methods: 3D print a network of channels within the polyurethane matrix using a sacrificial material like PVA. Dissolve the PVA post-formation to leave hollow channels and inject cyanoacrylate. Apply the final coating using spray-painting techniques.
Reasoning: Polyurethane provides durability while cyanoacrylate offers a rapid water-activated healing mechanism. The vascular-like structure allows multiple healing events, and the method is compatible with scalable techniques like spray-painting.

Evaluation:
Alignment with Research Objectives and Constraints: Score: 3 (Moderately Aligned)
Reasoning: The hypothesis addresses the objective of creating a self-healing coating with water-activated mechanisms. However, incorporating vascular-like channels via 3D printing may not be fully compatible with spray-painting techniques, which is a specified constraint. The use of 3D printing and sacrificial materials could complicate scalability and industrial application.

Scientific Plausibility: Score: 4 (Highly Plausible)
Reasoning: The use of polyurethane and cyanoacrylate is scientifically sound, and the concept of vascular networks for self-healing is established in materials science. Minor challenges may exist in integrating 3D printed channels with spray-painting, but overall, the hypothesis is plausible.

Innovation and Novelty: Score: 4 (Highly Innovative)
Reasoning: Combining vascular networks with spray-paintable coatings introduces a novel approach to self-healing materials. The integration of 3D printing techniques in creating self-healing coatings is a significant advancement.

Testability: Score: 2 (Difficult to Test)
Reasoning: Testing this hypothesis requires specialized equipment like 3D printers capable of creating vascular networks within polymers and handling sacrificial materials. This complexity makes experimental validation more challenging and resource-intensive.

Feasibility and Scalability: Score: 2 (Slightly Feasible)
Reasoning: Implementing vascular networks within coatings on an industrial scale is challenging. The process involves complex steps that are not easily scalable with current spray-painting technologies. Scaling up would require significant adjustments and investment.

Impact Potential: Score: 4 (High Impact)
Reasoning: If successfully implemented, this approach could significantly advance self-healing coatings by enabling multiple healing events and improving longevity, addressing critical challenges in corrosion protection.

Suggestion 2
Materials: Polyurea-based elastomer with polyisocyanate-filled vascular-like channels.
Methods: Use 3D printing to create a network of channels within the polyurea matrix using a sacrificial material. Dissolve the sacrificial material to leave the channels, then inject polyisocyanate. Apply using spray-painting techniques.
Reasoning: Polyurea and polyisocyanate are effective for corrosion protection. The vascular-like structure ensures multiple healing events, and the process is scalable.

Evaluation:
Alignment with Research Objectives and Constraints: Score: 3 (Moderately Aligned)
⋮
==
Ranking from Highest to Lowest Rated
Based on the evaluations, the suggestions can be ranked according to their overall scores:
Suggestions 4, 5, 6, 7, 9, 11, 13, 15, 17, 18, 19, 20 (Tied for Highest Rating)
Reasoning: These suggestions scored consistently high across all criteria, fully aligning with research objectives and constraints, being scientifically plausible, highly testable, completely feasible, and having high impact potential.

Suggestions 10 and 16 (Next Highest Rating)
Reasoning: These scored slightly lower in testability and feasibility due to minor challenges with fluoropolymer processing but still rated highly overall.

Suggestions 1, 2, 3, 8, 12, 14 (Lower Rating)
Reasoning: These suggestions, while innovative, faced challenges in alignment due to the use of vascular-like channels which complicate compatibility with spray-painting techniques, resulting in lower scores for testability and feasibility.

G Closeness Evaluation with BERTScore

To evaluate the similarity between the generated hypotheses and the ground truth hypotheses, we computed BERTScore Zhang et al. (2019) with the MaterialsBERT model (Shetty et al., 2023). The calculated scores for the proposed methods are as follows:

- **Hypotheses Generation Without Feedback:** 52.66%
- **Hypotheses Generation With Feedback:** 60.59%
- **Hypotheses Generation With Knowledge Graph and Feedback:** 50.30%

However, these scores do not align with the evaluations conducted by LLM-based methods or human experts in our study where Hypotheses Generation With Knowledge Graph and Feedback method achieves the best scores. Moreover, interpreting these scores within the context of our task poses challenges. For example, while there is a 7.93% improvement between the first two methods, it is unclear how this translates into factors such as the alignment of the mentioned properties, or the similarity of the methods described. This lack of interpretability makes BERTScore less effective for meaningfully assessing the performance of our methods in this domain.

Evaluation Metric	1	2	3	4	5
Alignment with Research Objectives and Constraints	Misaligned: Hypothesis does not relate to objectives and violates constraints.	Slightly Aligned: Hypothesis touches on the topic but misses key objectives or violates some constraints.	Moderately Aligned: Addresses some objectives but lacks focus on critical elements; meets most constraints.	Highly Aligned: Addresses most objectives and complies with major constraints.	Fully Aligned: Directly addresses all key objectives and fully adheres to all constraints.
Scientific Plausibility	Not Plausible: Contradicts fundamental scientific principles.	Slightly Plausible: Contains significant inaccuracies or unsupported assumptions.	Moderately Plausible: Scientifically sound but includes minor inaccuracies or speculative elements.	Highly Plausible: Scientifically accurate with negligible issues.	Completely Plausible: Fully consistent with established knowledge and well-supported by theory.
Innovation and Novelty	Not Innovative: Restates existing knowledge without introducing new ideas.	Slightly Innovative: Provides minimal new insights or slight variations on known concepts.	Moderately Innovative: Offers some original ideas or novel combinations of existing concepts.	Highly Innovative: Introduces significant new ideas or approaches.	Exceptionally Innovative: Groundbreaking ideas or methodologies that could revolutionize the field.
Testability	Not Testable: Cannot be tested with current techniques.	Difficult to Test: Requires specialized or prohibitively expensive equipment.	Moderately Testable: Can be tested with available techniques but requires complex procedures.	Easily Testable: Testable using common equipment and straightforward procedures.	Highly Testable: Rapid, cost-effective validation with readily available resources and techniques.
Feasibility and Scalability	Not Feasible: Cannot be implemented with current technology or resources.	Slightly Feasible: Theoretically possible but requires breakthroughs or high investment.	Moderately Feasible: Can be implemented on a small scale, but larger-scale production faces challenges.	Highly Feasible: Practical for small and moderate scales with manageable adjustments for scale-up.	Completely Feasible: Easily implemented and scaled from lab to industry with minimal changes or investment.
Impact Potential	No Impact: Unlikely to contribute meaningfully to the field.	Low Impact: Offers minor contributions but lacks significant influence.	Moderate Impact: Can contribute positively in specific areas.	High Impact: Likely to drive significant advancements or address important challenges.	Transformative Impact: Capable of revolutionizing the field or providing solutions to major problems.

Table 6: Evaluation metrics and scales for assessing generated hypotheses based on alignment with research objectives, scientific plausibility, innovation, testability, feasibility, and impact potential.