Interpreting LLM-Generated Scientific Hypotheses: Framework and Initial Findings

Adnan Mahmud*

Summary

This study investigates how LLMs generate scientific hypotheses by analyzing their interaction with research paper abstracts. Using TinyLlama-1.1B-Chat and attribution analysis, the research examined how different sections of scientific abstracts influence hypothesis generation. Key findings revealed that conclusion sections exert the strongest influence (23.7%), followed by methods (21.3%), while objectives showed the least impact (16.9%). The study also found no direct correlation between section length and attribution scores, suggesting content quality and positioning are more critical than length.

1. Introduction

In the thematic question of harnessing LLMs in the process of hypothesis generation, the remit of this report is to demonstrate a proof of concept and, to an extent, establish a framework for further research. To that end, this report examines how LLMs, given a set of abstracts from an arbitrary set of research papers, upon generate hypotheses and then analyses which sections of the abstracts have the most impact on the generated hypotheses, utilizing attribution analysis to quantify these relationships.

While LLMs demonstrate remarkable capabilities in processing and generating scientific text, there remains a critical gap in understanding how these models interpret and synthesize scientific information to generate novel hypotheses [1]. To that extent, this report focuses on analysing 5 abstracts from diverse scientific domains, examining how their structural components (background, objectives, methods, results, and conclusions) contribute to hypothesis generation. We employ a ***TinyLlama-1.1B-Chat***† model for hypothesis generation and develop a custom attribution analysis framework. The core questions this report asks are:

> [i.1] Which sections of scientific abstracts most strongly influence LLM hypothesis generation?
>
> [i.2] How consistent are these influences across different abstracts?
>
> [i.3] What patterns emerge in the relationship between section length and attribution scores?

* mam255@cantab.ac.uk

† TinyLlama-1.1B was chosen purely due to computational resource constraints, as 7B+ parameter models would require considerably more computational power. Furthermore, the Captum library is optimised for Meta-based products such as Llama.

The remainder of this report is organized as follows: Section 2 details our methodological framework and analysis pipeline. Section 3 presents our findings through quantitative analysis and visualizations. Section 4 discusses implications and limitations, and Section 5 concludes with recommendations for future research directions.

2. Methodology

Figure 1, a schematic representation of the hypothesis generation and attribution analysis experimental setup. The TinyLlama-1.1B-Chat language model (shown in green) receives input from both the scientific abstract and a human/user generated prompt. The model generates a hypothesis (shown in orange) based on these inputs. The Captum Feature Attribution framework (shown in diamond) performs a attribution analysis[‡] by examining the relationships between the generated hypothesis, the model's internal representations, and the original abstract. The framework analyzes the influence of different abstract sections on the hypothesis generation process (indicated by the green arrow returning to the abstract), providing quantitative measures of section-wise contributions.

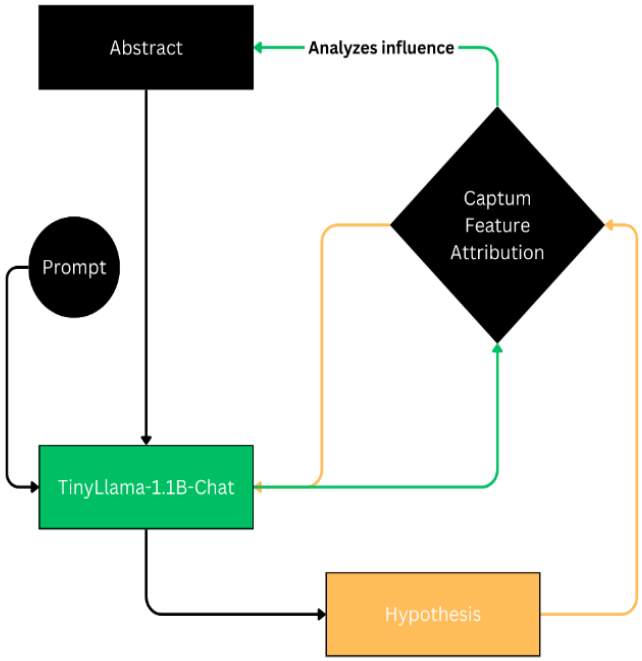

Figure 1: Experimental Architecture for Hypothesis Generation and Attribution Analysis.

2.1 Abstract Processing (Segmentation)

In transformer-based language models like TinyLlama, text is processed through tokenization, which breaks natural language into smaller units (tokens) for model processing. While tokenization is essential, it does not inherently preserve the structural information of natural language, which may affect human readability. To address this, we segment our abstract into five[§] sections and treat them as features: *Background*, *Objective*, *Methods*, *Results* and *Conclusion*.

[‡] a technique where a model's prediction is broken down to understand how much each input feature contributed to the final outcome

[§] Features could be defined in various ways. However, due to computational constraints, we heuristically chose to segment the abstract into five sections. One could define each word as a separate feature, but this would increase computational cost, especially in large-scale models.

Pseudo-algorithm 1: Abstract Segmentation Framework

```
Input: Scientific Abstract Text
PRE-PROCESSING:
  - Split abstract into sentences
  - Convert to lowercase for comparison
SECTION_CLASSIFICATION:
  For each sentence:
    IF contains section markers THEN
      Classify using Rule-Based Method
    ELSE
      Classify using Position-Based Method
      RULE-BASED_METHOD:
        - Check for section-specific keywords
        - Assign to corresponding section if match found
      POSITION-BASED_METHOD:
        IF abstract length ≤ 3 sentences THEN
          Assign all to methods
        ELSE
          Classify based on relative position:
          - First 20% → background
          - 20-40%  → objective
          - 40-60%  → methods
          - 60-80%  → results
          - Last 20% → conclusion
Output: Segmented Sections {background, objective, methods, results, conclusion}
```

2.2 Important Metrics

2.2.1 Attribution scores: quantify the degree to which each section of an abstract influences the model's hypothesis generation. Using the Feature Ablation[**] method implemented through the Captum framework, we systematically measure how changes in each section affect the model's output probability distribution.

For a given section ***s*** in abstract ***a***, the attribution score ***As,a*** is defined as:

$$A_{s,a} = P(h|a) - P(h|a \setminus s)$$

where:

P(h|a) is the probability of generating hypothesis h given the complete abstract a. This corresponds to the model's forward pass with complete input ***(forward_func(inputs))***.

P(h|a\s) is the probability of generating the same hypothesis when section s is ablated. Corresponding to the model's output when features are ablated ***(forward_func(ablated_inputs))***.

[**] Feature ablation is a perturbation-based attribution method where input features are replaced with a baseline to measure their impact on the output. By default, each scalar in a tensor is ablated independently, but a feature mask can group multiple features for collective ablation. This is useful in cases like image analysis, where entire segments can be tested for significance. When the model's output is fixed in size, perturbations per evaluation must be one, and any applied feature mask must be consistent across all inputs.

> The difference between these represents the attribution score, exactly matching Captum's implementation which they describe as "computing attribution, involving replacing each input feature with a given baseline / reference, and computing the difference in output" [3].

Higher attribution scores indicate sections that more strongly influence the generated hypothesis. A score of 0 suggests no influence, while higher scores (approaching 100) indicate sections critical to the hypothesis generation process.

2.2.2 Attribution Stability Metrics: These metrics evaluate the consistency of attribution scores across different abstracts, using measures like the Coefficient of Variation (CV) and Interquartile Range (IQR) to quantify how reliably each section influences hypothesis generation.

3. Results and Discussion

3.1 Primary Influences on LLM Hypothesis Generation [i.1]

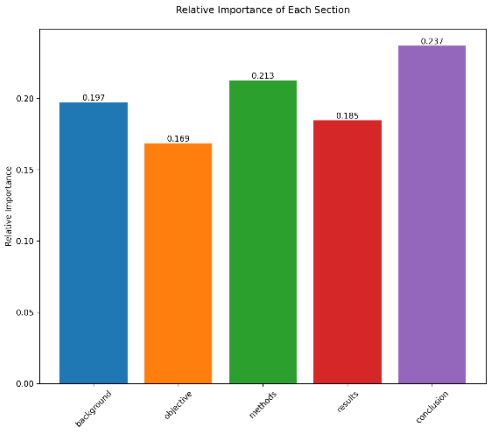

Figure 2: Relative Importance of Each Section. The proportional contribution of each section to hypothesis generation, normalized to sum to 1.0. This visualization directly addresses research question [I.1] by quantifying the relative influence of each section.

The analysis identifies a hierarchical structure in the influence exerted by abstract sections. As illustrated in *Figure 2*, the distribution of relative importance indicates that conclusion sections exert the strongest influence (23.7% of total attribution), followed closely by methods sections (21.3%). Background sections contribute significantly (19.7%), while results sections display moderate influence (18.5%). Objective sections, although least influential, maintain a notable impact (16.9%).

This hierarchy is further corroborated by the mean attribution scores (*Figure 3*), where conclusions demonstrate the highest mean score (77.85), followed by methods (68.30). Correlation analysis (Figure 1) reveals strong interrelationships between sections: a strong negative correlation (-0.95) between objectives and conclusions, a strong positive correlation (0.89) between results and objectives, and a moderate negative correlation (-0.51) between background and methods.

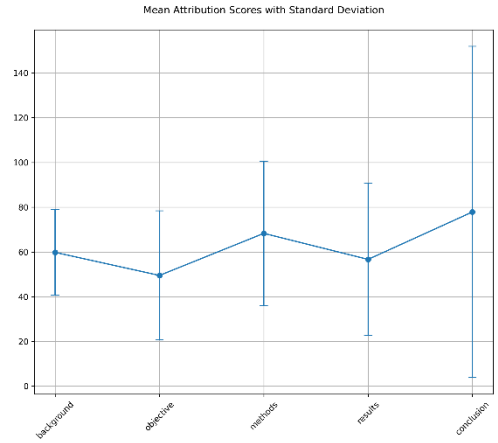

Figure 3: Mean Attribution Scores with Standard Deviation. A line plot showing the mean attribution scores (blue dots) with standard deviation error bars for each section. This visualization highlights both the average influence of each section and its variability, supporting research question [I.1] about which sections most strongly influence hypothesis generation. The large error bars for conclusions indicate high variability in their influence.

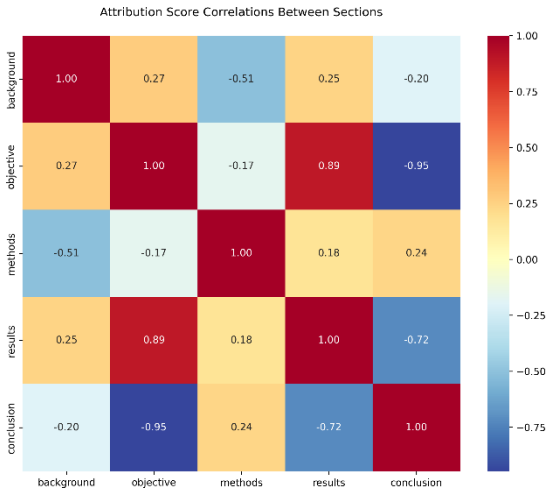

Figure 4: Attribution Score Correlations Between Sections. A heatmap visualization showing the correlation coefficients between different abstract sections' attribution scores. The color scale ranges from dark blue (-0.95) to dark red (1.00), with stronger correlations shown in more saturated colors. Notable relationships include the strong negative correlation between objectives and conclusions (-0.95) and the strong positive correlation between results and objectives (0.89). This visualization directly addresses research question [I.2] by showing how different sections' influences relate to each other across abstracts.

3.2 Consistency of Sectional Influence [i.2]

The stability of these influences varies across abstracts, as indicated by multiple stability metrics (Figures 5 and 6). Background sections exhibit the highest consistency (CV: 0.32), while methods sections display moderate stability (CV: 0.47). Conversely, conclusions demonstrate the greatest variability (CV: 0.95, IQR: 45.45), with results sections also exhibiting significant variation (CV: 0.60), and objectives showing notable fluctuations (CV: 0.58).

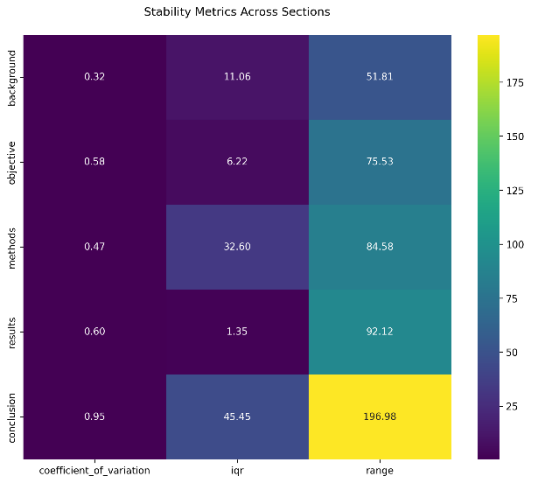

Figure 5: Stability Metrics Across Sections. A heatmap showing three stability metrics (coefficient of variation, IQR, and range) for each section. This visualization supports research question [I.2] by providing multiple measures of influence consistency.

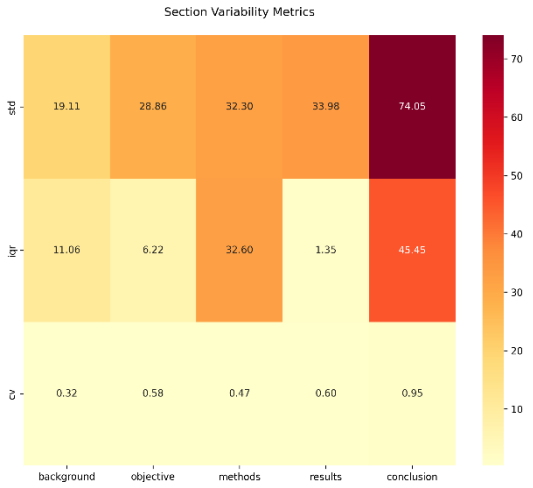

Figure 6: Section Variability Metrics. A heatmap focusing on three key variability metrics (standard deviation, interquartile range, and coefficient of variation) across sections.

The violin plot (Figure 7) visualises these distribution patterns, highlighting the widest distribution for conclusion sections and more concentrated distributions for background sections. These findings suggest that while conclusions are generally the most influential, their impact varies considerably across different abstracts.

3.3 Section Length and Attribution Relationships [i.3]

An analysis of the relationship between section length and attribution scores (Figure 8) reveals several key patterns. No strong linear correlation is observed, indicating that section length does not directly determine attribution scores. Short sections, particularly conclusions, can achieve high attribution scores, while longer sections do not necessarily yield greater influence.

Efficiency patterns emerge, with conclusions attaining high attribution scores (up to 196.98) despite their relatively short length. Methods sections exhibit consistent attribution scores across varying lengths, while background sections maintain a stable influence regardless of length. The majority of sections cluster between 100-400 characters, yet attribution scores vary widely within similar length ranges, with medium-length sections often producing the highest attribution scores. These findings suggest that content quality and positioning, rather than section length, are critical determinants of influence in hypothesis generation.

3.4 Limitations of the experiments

The segmentation process has many errors [Appendix 2]. The core limitation of this experiment, beyond the obvious one of low raw data (only five abstracts), is that it is entirely attribution-based. **Thus, even in the best-case scenario, one can merely make inferences about the LLM's decision process. A mechanistic interpretability would be a potential cure [4].**

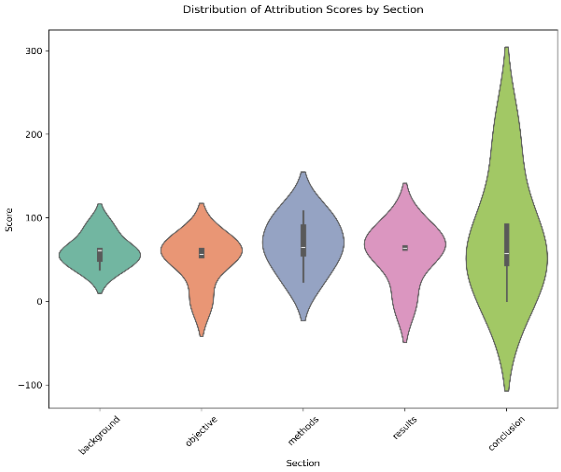

Figure 7: Distribution of Attribution Scores by Section. A violin plot showing the probability density of attribution scores for each abstract section. The width of each "violin" represents the frequency of scores at that level, while the internal box plots show.

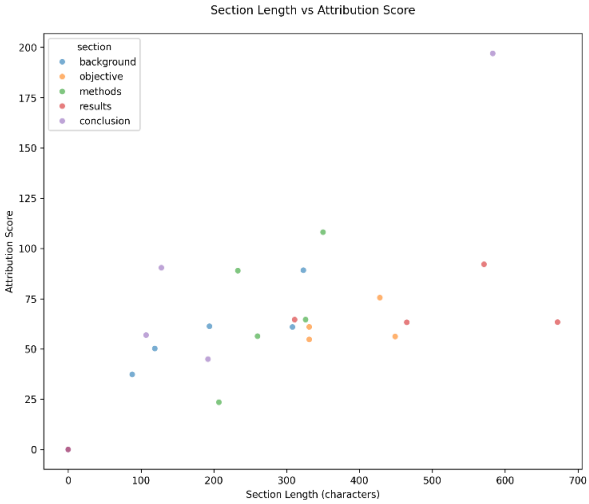

Figure 8: Section Length vs Attribution Score. A scatter plot showing the relationship between section length (x-axis, measured in characters) and attribution scores (y-axis). Each point represents a section from an abstract, color-coded by section type. The plot reveals no strong linear relationship between length and attribution score, with high scores possible at various lengths. This visualization directly addresses research question [I.3] by illustrating how section length relates to influence on hypothesis generation.

4. Broader Research Agenda

To contextualise the observations of this report within the overarching *thematic*[††] and *specific*[‡‡] research questions, our findings indicate that LLM hallucination in hypothesis generation follows a structured pattern rather than occurring randomly. The clear hierarchy of sectional influences, with conclusions (23.7%) and methods (21.3%) exerting the greatest impact, suggests that the model's "creative" outputs are anchored in particular aspects of the input text. This structured influence pattern implies that LLM hallucinations in scientific contexts resemble guided

[††] In the domain of scientific discovery, how can we understand and potentially harness LLMs' hallucination mechanisms, for "creative" hypothesis generation?
[‡‡] How do LLMs Select Between Multiple Hypotheses?

extrapolation rather than arbitrary generation. The strong negative correlation (-0.95) between objectives and conclusions is particularly significant, as it suggests that the model actively balances stated research aims against reported findings when formulating new hypotheses. This compensatory mechanism may represent a form of "controlled hallucination," wherein the model generates novel hypotheses by bridging gaps between research objectives and conclusions. Furthermore, the relationship between section length and attribution scores suggests a sophisticated mechanism for balancing creative freedom with scientific rigour. The observation that medium-length sections frequently yield the highest attribution scores indicates an optimal information density for hypothesis generation—providing sufficient detail to ensure scientific validity without excessively constraining creative extrapolation. This suggests that LLMs utilise conclusions as primary sources for creative hypothesis generation while maintaining links to methodological and background content to uphold scientific plausibility.

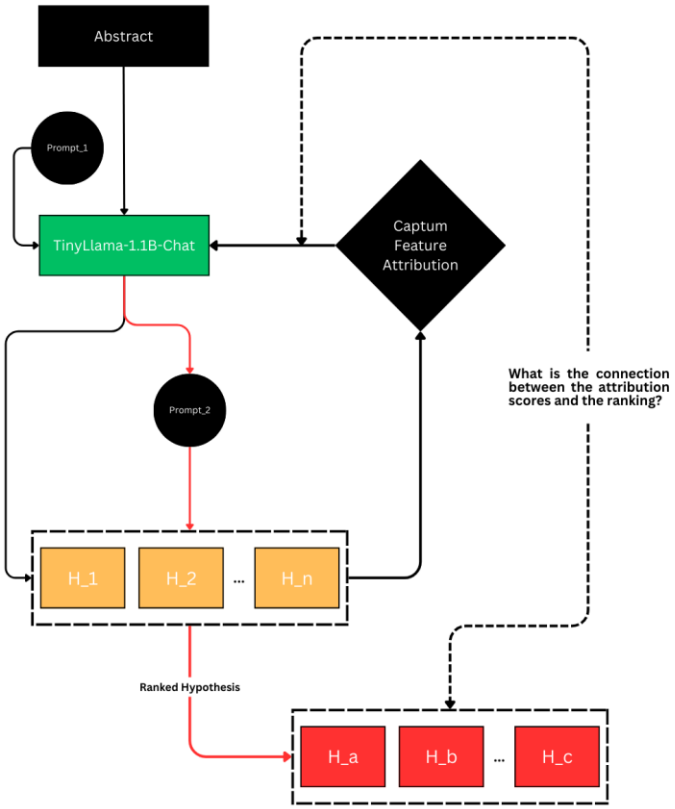

Figure 9: Potential experimental Architecture for the next report.

The differing consistency patterns across sections provide further insight into how LLMs navigate between multiple potential hypotheses. The high consistency of background sections (CV: 0.32) suggests they function as stabilising elements in hypothesis selection, potentially filtering out scientifically implausible hallucinations. Whereas the high variability in conclusion sections (CV: 0.95) indicates that this is where the model has the greatest freedom for creative hypothesis generation, with conclusions serving as departure points for novel ideas.

5. Next Experimental Setup

This diagram illustrates the experimental setup for investigating the relationship between hypothesis generation, ranking, and attribution analysis. The process begins with an *abstract* and initial prompt (*Prompt*$_1$) being fed into *TinyLlama-1.1B-Chat* (shown in green), which generates multiple hypotheses (H_1 through H_n, shown in orange). A second prompt (*Prompt*$_2$) [Appendix 3] is used to have the model rank these hypotheses. The ranked hypotheses (H_a through H_c, shown in red) are then analyzed using the *Captum Feature Attribution framework* (black diamond) to understand the connection between attribution scores and the model's ranking decisions. The

black arrows represent the initial hypothesis generation flow, while red arrows show the ranking process flow. The dashed feedback loop indicates the analysis of how attribution patterns correlate with hypothesis rankings.

The key research question, highlighted in the diagram, is "What is the connection between the attribution scores and the ranking?" This investigation aims to understand how the model's internal attention patterns and section influences relate to its preferences when ranking hypotheses.

Conclusion

The findings indicate that LLM hypothesis generation follows structured patterns rather than random hallucination. The clear hierarchical influence of abstract sections, particularly the strong negative correlation between objectives and conclusions (-0.95), suggests that LLMs employ a form of "controlled hallucination" - bridging gaps between research objectives and findings when generating novel hypotheses. The model uses conclusions as primary sources for creative generation while maintaining links to methodological and background content for scientific plausibility. This structured approach implies that LLM hallucinations in scientific contexts represent guided extrapolation rather than arbitrary generation.

References

[1] Cohrs, K.-H., Diaz, E., Sitokonstantinou, V., Varando, G., & Camps-Valls, G. (2025). Large language models for causal hypothesis generation in science. Machine Learning: Science and Technology, 6(1), 013001. https://doi.org/10.1088/2632-2153/ada47f

[2] Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbering, M., Leveling, J., ... & Flores-Herr, N. (2024). Tokenizer Choice For LLM Training: Negligible or Crucial? arXiv preprint arXiv:2310.08754v4. https://doi.org/10.48550/arXiv.2310.08754

[3] Captum. (n.d.). Feature Ablation. Retrieved from https://captum.ai/api/feature_ablation.html

[4] Bereska, L., & Gavves, E. (2024). Mechanistic Interpretability for AI Safety -- A Review. *arXiv preprint arXiv:2404.14082v3*. https://doi.org/10.48550/arXiv.2404.14082

Appendices

Appendix 1: Sentence Distribution Per Segments

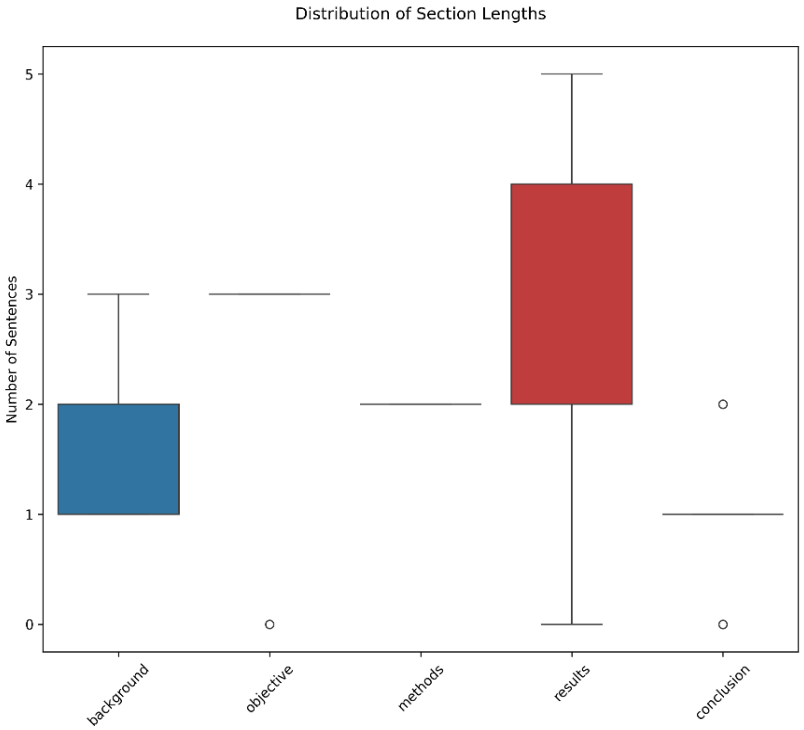

Figure 10: Distribution of sentence lengths across five segments of five scientific abstracts, showing the variation in section sizes using box plots.

The box plot visualization demonstrates the typical distribution of sentences across different abstract sections in our dataset. Results sections show the highest variability, containing between 0 to 5 sentences, with a median around 3 sentences. Background sections typically contain 1-2 sentences, while methods sections show more consistency with most papers containing 2 sentences. Notably, objective and conclusion sections tend to be the most concise, often containing just one sentence, with some abstracts lacking explicit objective statements (shown by outliers at 0).

Appendix 2: The paper title, paper abstract, prompt and the LLM generated hypothesis in a tabular form. The color mask is as follows: {background:yellow}, {objective:green}, {methods:blue}, {conclusion:red}.

Paper Title	Paper Abstract	Prompt	LLM Hypothesis
Automating the practice of science: Opportunities, challenges, and implications.	Automation transformed various aspects of our human civilization, revolutionizing industries and streamlining processes. In the domain of scientific inquiry, automated approaches emerged as powerful tools, holding promise for accelerating discovery, enhancing reproducibility, and overcoming the traditional impediments to scientific progress. This article evaluates the scope of automation within scientific practice and assesses recent approaches. Furthermore, it discusses different perspectives to the following questions: where do the greatest opportunities lie for automation in scientific practice?; What are the current bottlenecks of automating scientific practice?; and What are significant ethical and practical consequences of automating scientific practice? By discussing the motivations behind automated science, analyzing the hurdles encountered, and examining its implications, this article invites researchers, policymakers, and stakeholders to navigate the rapidly evolving frontier of automated scientific practice.	Read this scientific paper abstract and identify its main hypothesis.	The main hypothesis proposed by this scientific article is that automation holds immense potential in transforming the practice and outcomes of scientific research. The paper explores different perspecitves on what automation can offer, including new opportunities for speed and efficiency, enhancement of reproducibilty, and addressing various cultural barriers. While there are some limitations and obstacles to be addressed when implementing such technologies, the study suggests that automating science could significantly benefit society's knowledge generation process
Random Forests for Heteroscedastic Data	Random forests are a popular machine learning technique that are effective across a range of scientific problems. We extend the standard algorithm to incorporate the uncertainty information that arises in heteroscedastic data - datasets where the amount of noise in the target value varies between datapoints. We consider datasets where the relative amount of measurement noise in different datapoints is known. This is not the standard scenario, but does commonly exist in real data, as we illustrate on 10 drug design datasets. Utilising this uncertainty information can lead to significantly better predictive performance. We introduce three random forest variations to learn from heteroscedastic data: parametric bootstrapping, weighted random forests and variable output smearing. All three can improve model performance, demonstrating the adaptability of random forests to heteroscedastic data and thus expanding their applicability.	Read this scientific paper abstract and identify its main hypothesis.	The main hypothesis proposed by this scientific article is to extend the traditional Random Forest (RF) algorithm to include uncertainty information about heteroskedasticity in data. The authors propose three variations of RF, each with some specific features designed to address this phenomenon, namely parametric bootstrap, weight-boostrap, and variable smear. By comparing these approaches' performances on several drug discovery datasets, it becomes evident that all the three differ in their contribution towards improving model accuracy while still keeping stability. Therefore, the author concludes that all variants have proven their usefulness in handling heteroshedastic data.

	Additionally, variations in the relative performance of the three methods across datasets provides insight into the mechanisms of random forests and the purpose of the different random elements within the model.		
Genesis: Towards the Automation of Systems Biology Research	The cutting edge of applying AI to science is the closed-loop automation of scientific research: robot scientists. We have previously developed two robot scientists: `Adam' (for yeast functional biology), and `Eve' (for early-stage drug design)). We are now developing a next generation robot scientist Genesis. With Genesis we aim to demonstrate that an area of science can be investigated using robot scientists unambiguously faster, and at lower cost, than with human scientists. Here we report progress on the Genesis project. Genesis is designed to automatically improve system biology models with thousands of interacting causal components. When complete Genesis will be able to initiate and execute in parallel one thousand hypothesis-led closed-loop cycles of experiment per-day. Here we describe the core Genesis hardware: the one thousand computer-controlled - bioreactors. For the integrated Mass Spectrometry platform we have developed AutonoMS, a system to automatically run, process, and analyse high-throughput experiments. We have also developed Genesis-DB, a database system designed to enable software agents access to large quantities of structured domain information. We have developed RIMBO (Revisions for Improvements of Models in Biology Ontology) to describe the planned hundreds of thousands of changes to the models. We have demonstrated the utility of this infrastructure by developed two relational learning bioinformatic projects. Finally, we describe LGEM+ a relational learning system for the automated abductive improvement of genome-scale metabolic models.	Read this scientific paper abstract and identify its main hypothesis.	The main hypothesis behind this scientific article is that the use of robot scientism, i.e., automatic investigation of complex systems using robots as tools instead of humans, could significantly reduce costs and increase efficiency when it comes to understanding biological processes. Specifically, this study explores the concept of \"closed-loop\" automation where robots self-learn from their mistakes while simultaneously increasing the speed and accuracy of experimental outcomes compared to traditional laboratory techniques. This approach has the potential to lead to significant advancements in areas such as pharmaceutical discovery, cancer therapy, and food safety. However, further development and refinement of these robots is still needed before they can fully replace experienced scientists in some fields.
Extrapolation is not the same as interpolation	We propose a new machine learning formulation designed specifically for extrapolation. The textbook way to apply machine learning to drug design is to learn a univariate function that when a drug (structure) is input, the function outputs a real number (the activity): f(drug) activity. However, experience in real-world drug design suggests that this		The main hypothesis proposed in this scientific article is that the traditional approach to drug development using regression analysis may miss some relevant information beyond the range covered by the training data, particularly for extracurrent

	formulation of the drug design problem is not quite correct. Specifically, what one is really interested in is extrapolation: predicting the activity of new drugs with higher activity than any existing ones. Our new formulation for extrapolation is based on learning a bivariate function that predicts the difference in activities of two drugs F(drug1, drug2) difference in activity, followed by the use of ranking algorithms. This formulation is general and agnostic, suitable for finding samples with target values beyond the target value range of the training set. We applied the formulation to work with support vector machines , random forests , and Gradient Boosting Machines . We compared the formulation with standard regression on thousands of drug design datasets, gene expression datasets and material property datasets. The test set extrapolation metric was the identification of examples with greater values than the training set, and top-performing examples (within the top 10% of the whole dataset). On this metric our pairwise formulation vastly outperformed standard regression. Its proposed variations also showed a consistent outperformance. Its application in the stock selection problem further confirmed the advantage of this pairwise formulation.	Read this scientific paper abstract and identify its main hypothesis.	prediction tasks where the sample values are outside the target range but still have high levels of intervention. The author proposes a novel formulation called \"extrapolation\" which uses a bipartite graph to represent relationships between different pairs of targeted drugs and non-targeted drugtarget combinations. This approach can be generalized to find samples beyond the training boundaries while taking into account relevance to the target area. The experimental results show that the proposed method performs consistently better than standard regression methods, showing potential applications in several fields such as stock selection and genome sequencing pipelines.
Scientific Hypothesis Generation by a Large Language Model: Laboratory Validation in Breast Cancer Treatment	Large language models (LLMs) have transformed AI and achieved breakthrough performance on a wide range of tasks that require human intelligence. In science, perhaps the most interesting application of LLMs is for hypothesis formation. A feature of LLMs, which results from their probabilistic structure, is that the output text is not necessarily a valid inference from the training text. These are 'hallucinations', and are a serious problem in many applications. However, in science, hallucinations may be useful: they are novel hypotheses whose validity may be tested by laboratory experiments. Here we experimentally test the use of LLMs as a source of scientific hypotheses using the domain of breast cancer treatment. We applied the LLM GPT4 to hypothesize novel pairs of FDA-approved non-cancer drugs that target the MCF7 breast cancer cell line relative to the non-tumorigenic breast cell line MCF10A. In the first round of	Read this scientific paper abstract and identify its main hypothesis.	The main hypothesis proposed in the paper is that large language models can provide a valuable resource for generating novel scientific hypothoses through their ability to form hallucination hypotheses based on their probabilistically structured outputs. The study applies a large language model (GPT4) to discover novel drug pairings related to breast cancer therapy, and conducts further experiments with similar results. While GPT's explanations lack conviction, the use case demonstrated the potential of ML as a tool for hypothesis generation within scientific domains such as medicine.

	laboratory experiments GPT4 succeeded in discovering three drug combinations (out of 12 tested) with synergy scores above the positive controls. These combinations were itraconazole + atenolol, disulfiram + simvastatin and dipyridamole + mebendazole. GPT4 was then asked to generate new combinations after considering its initial results. It then discovered three more combinations with positive synergy scores (out of four tested), these were disulfiram + fulvestrant, mebendazole + quinacrine and disulfiram + quinacrine. A limitation of GPT4 as a generator of hypotheses was that its explanations for them were formulaic and unconvincing. We conclude that LLMs are an exciting novel source of scientific hypotheses.		

Appendix 3: Potential Prompt$_2$

Here are several hypotheses generated from the scientific abstract:
[H_1]: [First hypothesis text]
[H_2]: [Second hypothesis text]
...
[H_n]: [Nth hypothesis text]
Rank these hypotheses from best to worst, with brief explanations for your choices.