advances in Technology

Adv. Technol. 2026, 6(1), 1-16

Full Paper

Connection between Temperature and Emergent Behavior in Large Language Models

Nipun D. Balage, Lahiru D. Mohottalage, Ashen S. Wickramasinghe, M.P.P. Liyanage*, and H.M.S.C.R. Heenkenda

Department of Information and Communication Technology, Faculty of Technology, University of Sri Jayewardenepura, Homagama 10206, Sri Lanka.

Corresponding Author: prabani@sjp.ac.lk

Received: 20 December 2025; Revised: 09 February 2026; Accepted: 29 April 2026; Published: 22 August 2026

Abstract

This paper investigates the effect of the Sampling Temperature Parameter on behavioral changes, known as Emergent Behavior, in Large Language Models (LLM). Emergent Behavior was previously attributed to a direct result of an LLM's scale. This study explores whether the Sampling Temperature Parameter has any effect on Emergent Behavior with three LLMs of various sizes. These LLMs were tested using 100 questions that belong to 5 categories, with controlled experiments at 11 Sampling Temperature settings. The findings show that LLMs tend to ignore user prompts, changing their behavior with higher temperatures. These findings suggest that temperature changes strongly affect the LLM's behavior and induce abrupt changes, which can be attributed to emergent-like behavior. Even smaller language models exhibit emergent-like behaviors, suggesting that emergence is not solely a function of model size. This study presents a critical, underexplored area of research that is capable of triggering nonlinear performance changes. This paper broadens the theoretical understanding of LLMs' dynamics and offers practical value in controlling their behavior for safer and more reliable use.

Keywords: artificial intelligence, emergent behavior, generative artificial intelligence, large language models, sampling temperature

Introduction

One of the most fascinating aspects of LLMs is their ability to change their behavior in unpredictable ways, which is known as emergent behavior. Emergent behavior refers to the phenomenon in which abrupt and complex abilities arise as a model scales in terms of its size, training data, or computational resources. This phenomenon has been well documented in recent literature [1, 2].

Many of the previous studies attribute such emergent capabilities of models to factors like the model's scale and its architecture, but the role of the sampling temperature parameter, if it has any, remains underexplored. Temperature is one of those hyperparameters that is applied during inference, and it can be used to control the randomness of the sampling within an LLM. Sampling temperature is a modification to the softmax function, which is at the final layer of an LLM that is used to convert raw scores into probabilities, which is how it selects the next word that comes in the generated phrase [3].

Higher temperature leads to more creative responses, but they can be potentially erratic outputs. Lower temperatures output more consistent, deterministic, and rigid responses [4, 5].

However, the sampling temperature's potential to induce emergent behavior has not been rigorously investigated. This gap is significant. If the alterations in the temperature value can induce behavioral changes in an LLM, we can suggest that the model scale or architecture may not directly induce emergence. The knowledge is crucial in applications that require safety, consistency, and reliability in the applications of generative models. This offers novel channels of regulating the behavior of LLM without retraining them or altering their architecture, which can be expensive and time-consuming.

Related Work

Emergent capabilities and the role of inference parameters, such as temperature, have been thoroughly studied with the rapid advancement of LLMs. The previous studies done on this topic show how the scale of the model can trigger chain-of-thought reasoning that does not appear to be present in smaller models [1, 6]. Previous studies suggest that emergent behavior may arise from complex interactions among model architecture, training objectives, and the diversity of training data [6]. However, the precise mechanisms underlying how and why such emergent capabilities occur remain unclear [7].

The temperature parameter's effect on LLMs has also been extensively studied on various subjects. Lower temperature values that range from 0.1 to 0.5 result in very deterministic, highly probable responses. Lower temperature values are used for factual tasks that require distinct answers [6]. Higher temperature values that range from 0.7 to 1.2 and up increase creative outputs from a model at the cost of coherence [5]. Recent work on this topic also reveals that higher temperature values can introduce security vulnerabilities, making models more prone to adversarial attacks [8].

Previously, studies have explored how an LLM behaves when introduced with different temperatures [3, 9]. Renze et al., (2024) have reported how sampling temperature affects the performance of LLMs when faced with various problem-solving tasks, combined with how prompt engineering may improve these results. They have studied this using multiple LLMs and a questions and answers method similar to ours. In [9], the study also focuses on how LLMs accuracy changes with different temperatures by evaluating them with a data set designed to assess six capabilities of an LLM. Both of the studies focus on the performance of an LLM, while our study focuses on direct behavioral changes in an LLM by examining the deviations from our rules/prompts which leads to the LLM producing wrong results.

While both topics are thoroughly studied in their own scope, no existing work explicitly examines whether the two topics have any interaction. Emergent behavior is typically analyzed with a model scale [1], and temperature is studied as an inference time control parameter [4, 5]. Key questions remain unanswered: does temperature modulate the intensity of emergent behaviors (e.g., amplifying reasoning errors in larger models)?

Materials and Methods

In this study, we systematically evaluated the responses of LLaMA-v3.1-8B (accessed via the Fireworks AI platform), ChatGPT-4o-mini with OpenAI, and Claude-3.7-Sonnet with Anthropic across a controlled range of temperatures to determine emergent-like behavior. We selected three LLMs of various sizes to ensure robust comparative analysis. In these three models, LLaMA-v3.1-8B was chosen as the smaller model, while the other two models are more advanced models.

A total of 100 questions were curated and refined by the authors and stratified across five distinct categories to evaluate model performance: Mathematical Numerical, Text-Based Mathematical, Logical Questions, General Knowledge, and True/False.

Each category included 20 questions, designed to test both factual retrieval and reasoning ability under stochastic sampling conditions.

To analyze the relationship between temperature and emergent behavior, we tested each model at 11 discrete temperature values. For the Llama model, this spanned 0.0 to 1.0 in increments of 0.1. For GPT and Claude, which support a wider temperature range, we normalized their evaluation points across 0.0 to 2.0 in increments of 0.2, ensuring parity in the number of evaluation points and maintaining comparability across models.

Experiments were conducted using API-driven automated testing, ensuring identical prompts across all models throughout the entire study, response logging, and reproducibility by monitoring how the LLM behaves to the user prompt as the temperature increases. The prompts were designed in a way that leads the LLM to provide an invalid result if it chooses to ignore the user prompt. The complete source code along with the prompts and question set can be found in the Supporting Information section.

The LLM outputs were subsequently visualized through heat maps and line graphs to detect non-linear behavioral shifts indicative of emergent properties. Each question was tested 5 times, and the best-performing output was chosen for the visualization.

We used task-specific grading schemes, depending on the question category, to analyze the outputs across different categories. Both numerical studies, including Mathematical Numerical and Mathematical Text-Based questions, were scored based on the percentage error from the correct answers [8]. True/False responses were graded with exact matching because they only have binary answers. However, Logical Questions and General Knowledge questions needed a method that could handle more subjective outputs. We applied the rubric evaluation method to measure the correctness, completeness, and clarity of such responses [8]. This allowed for nuanced detection of the degree of deviation from user prompts, hallucinations or unexpected reasoning.

The use of standardized, stratified evaluation systems and strongly regulated inputs made it easy to isolate temperature as a single variable of interest to systematically observe the behavioral dynamics in LLMs. This methodological approach allowed comparative studies on varying model structures and temperatures, supporting the discovery of emergent-like behaviors.

Results and Discussion

Here, we provide the results of the study of the three LLMs, which are LLaMA-v3.1-8 B, ChatGPT-4o-mini, and Claude-3.7-Sonnet, grouped by question categories. The study utilized a strictly organized dataset and controlled temperature variants to examine the response patterns of each of the models in relation to the five question categories. The analysis mainly focuses on emergent-like behavior identification and changes in correctness, coherence, creativity, and the reasoning style. Each section of categories is presented with two types of visualizations. A heat map showing how each and individual question performed in each temperature setting and a line chart demonstrating each question's performance at a given temperature.

Mathematical Numerical

To provide a more detailed analysis of model behavior, the performance of individual Mathematical Numerical questions was examined across different temperature settings. Heat maps were used to visualize the correctness of responses generated by each model for every question at temperature values.

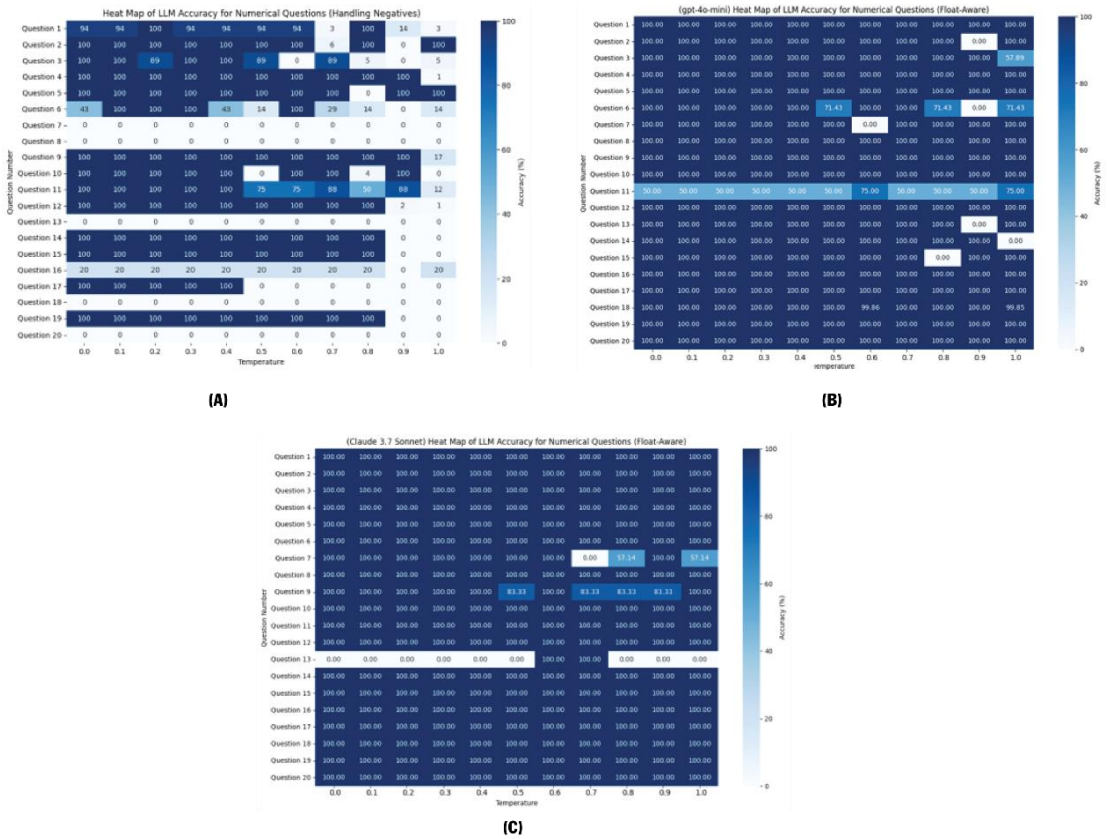

Figure 1. Mathematical Numerical questions performed at each temperature value. **(A)** Llama-v3.1-8b Heat map, **(B)** ChatGPT-4o-mini Heat map, and **(C)** Claude-3.7-sonnet Heat map

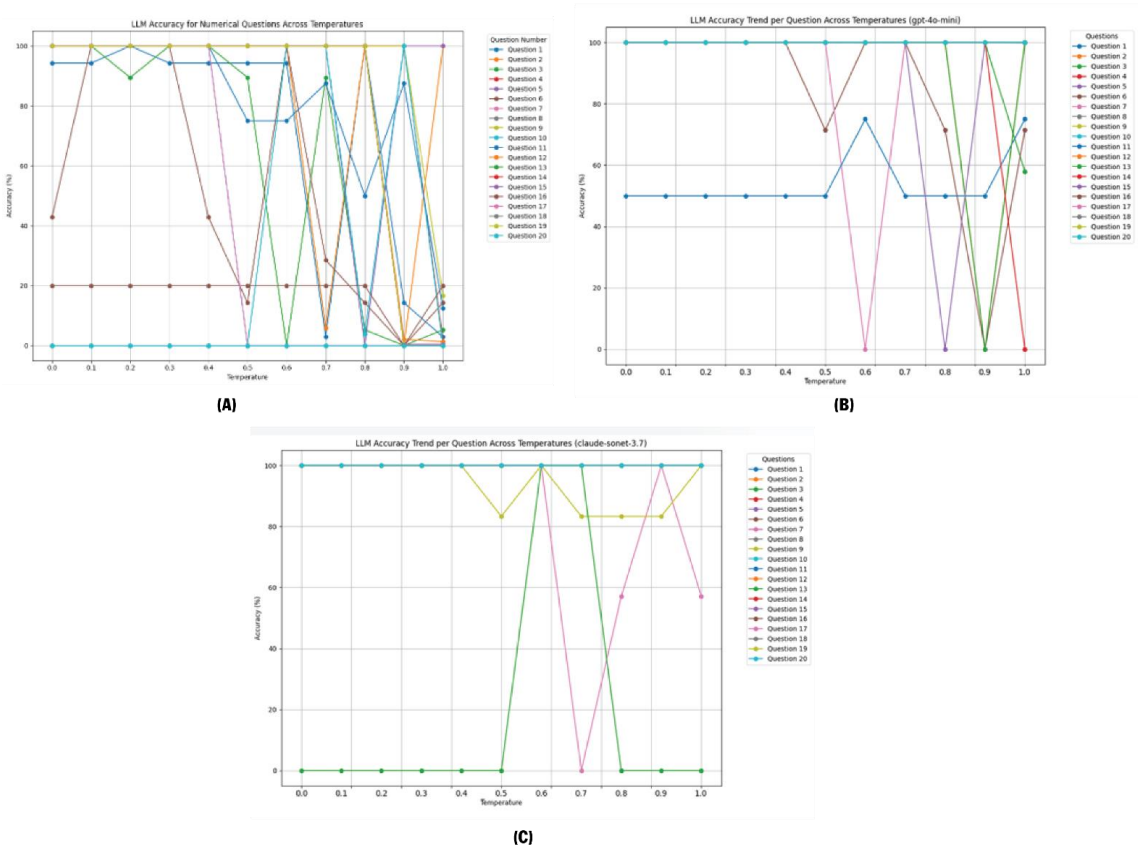

Figure 2. (A) Llama-v3.1-8b Line Chart, **(B)** ChatGPT-4o-mini Line Chart, and **(C)** Claude-3.7-sonnet Line Chart

Figures 1(A)-(C) heat maps show how individual mathematical numerical questions performed at each temperature value. All three models remained submitted to the user prompts consistently at lower temperature settings, which led to them producing more accurate answers, particularly between 0.0 and 0.5. This stability reflects the deterministic behavior of LLMs under minimal sampling randomness. However, as temperature increased, a clear degradation was observed, marked by abrupt accuracy drops, and erratic trends, especially as seen in Figure 1(A) and 1(B), which visualize the performance of LLaMA-v3.1-8b and ChatGPT-4o-mini, respectively. Claude-3.7-Sonnet, shown in Figure 1(C), while generally more robust, also exhibited instability at higher temperatures, though with smoother transitions in most cases.

This chaotic behavior in higher temperature can be clearly seen in the line charts in Figures 2(A)-(C). LLaMA-v3.1-8b and ChatGPT-4o-mini shown in Figure 2(A) and Figure 2(B) shows chaotic behavior in higher temperatures, this can also be seen in Figure 2(C) which visualizes Claude-3.7-Sonnet despite seeming to perform better in heatmap visualization.

Mathematical Text-Based

To further evaluate the effect of temperature on reasoning and problem-solving capabilities, the performance of the models on Text-Based Mathematical questions was analyzed.

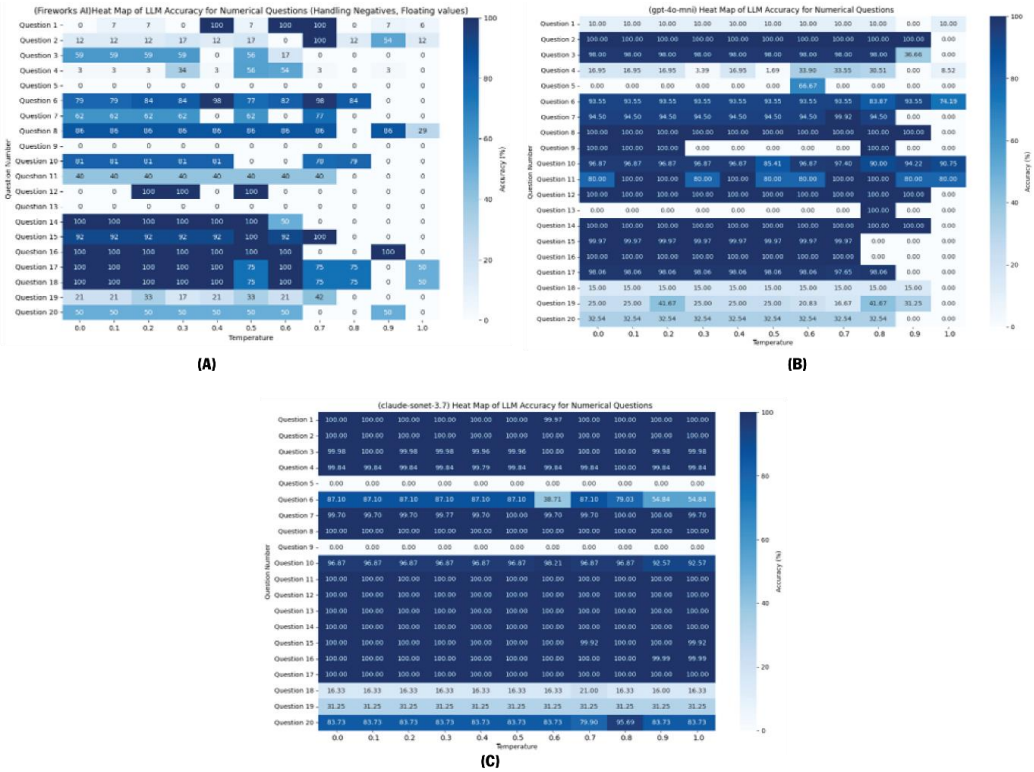

Figure 3. (A) Llama-v3.1-8b Heat map, **(B)** ChatGPT-4o-mini Heat map, and **(C)** Claude-3.7-sonnet Heat map

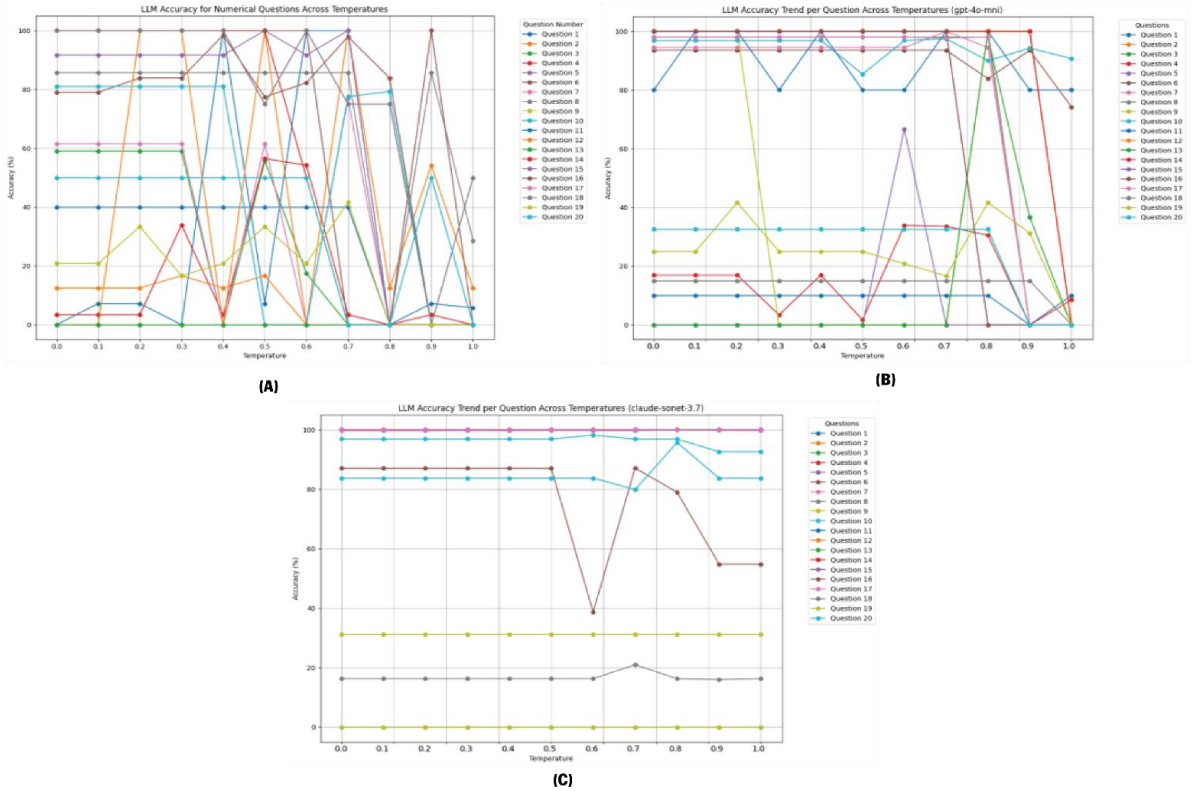

Figure 4. (A) Llama-v3.1-8b Line Chart, **(B)** ChatGPT-4o-mini Line Chart, and **(C)** Claude-3.7-sonnet Line Chart

The results presented in Figures 3(A) and 3(B) indicate that both LLaMA-v3.1-8B and ChatGPT-4o-mini achieved consistency at lower temperatures, demonstrating stability across most text-based mathematical questions. Claude-3.7-Sonnet, which is shown in Figure 3(C), outperformed the others overall, showing a uniform distribution of consistency across the temperature range, with only minor degradation at the highest settings.

However, as temperature increased, all three models exhibited volatility in their outputs. Consistency began to fluctuate sharply beyond mid-temperature levels, with question-level performance showing irregular drops and spikes. This can be seen in Figures 4(A) and Figure 4(B), which shows how LLaMA-v3.1-8B and ChatGPT performed overall. While Claude remained comparatively more stable, signs of emergent instability were still evident at higher temperatures, as seen in Figure 4(C), especially after the temperature point of 0.5.

True or False

The True/False category was included to evaluate the models' ability to make accurate binary decisions under varying temperature settings.

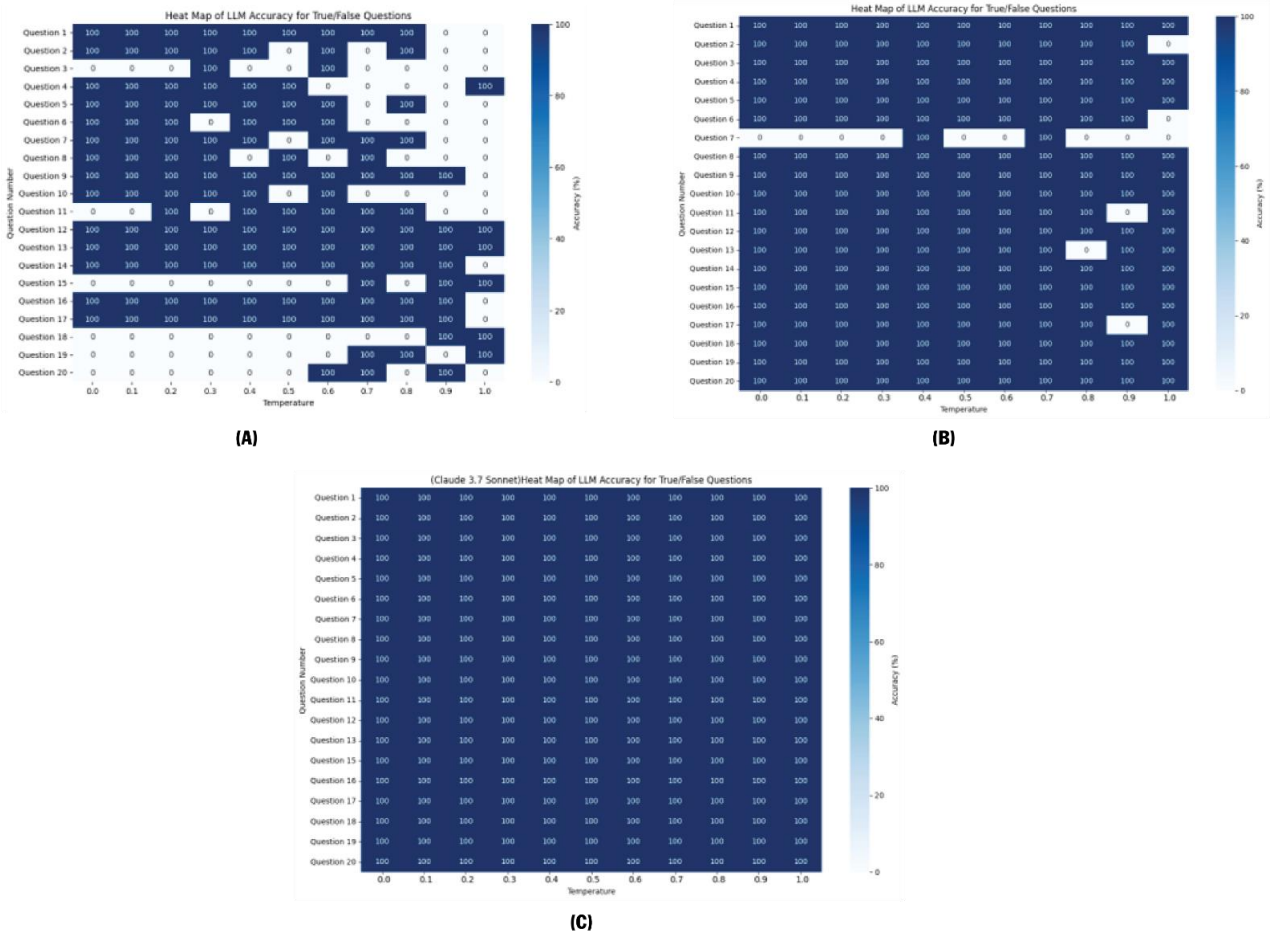

Figure 5. (A) Llama-v3.1-8b Heat map, **(B)** ChatGPT-4o-mini Heat map, and **(C)** Claude-3.7-sonnet Heat map

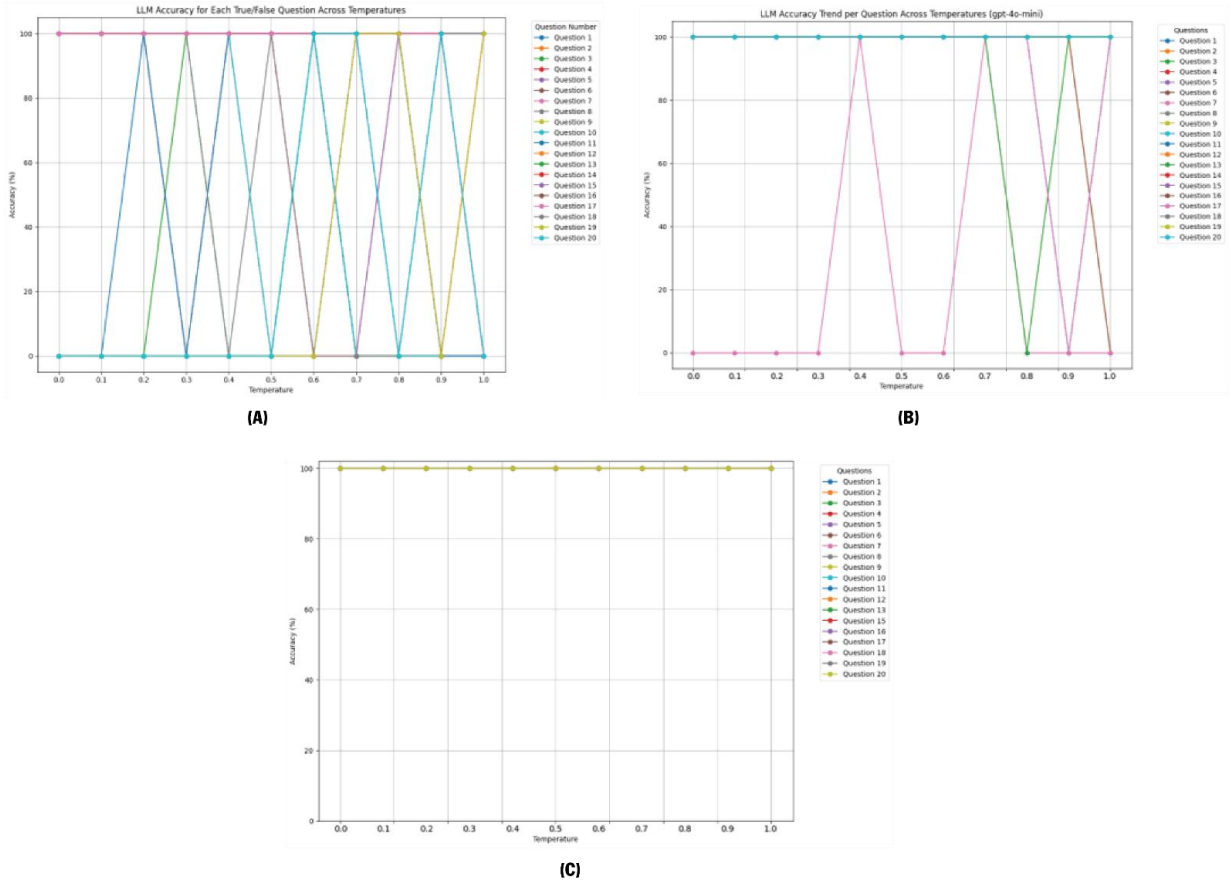

Figure 6. (A) Llama-v3.1-8b Line Chart, **(B)** ChatGPT-4o-mini Line Chart, and **(C)** Claude-3.7-sonnet Line Chart

As shown in Figures 5(A) and 5(B), both LLaMA-v3.1-8B and ChatGPT-4o-mini demonstrated consistent responses at lower temperature settings, indicating greater stability and determinism in their responses to the given prompts. As temperature increased, both models showed a marked decline in reliability, with noticeable drops and fluctuations across multiple questions which is most apparent in Figure 5(A) with LLaMA-v3.1-8B. In contrast, Figure 5(C) and Figure 6(C), which visualize the Claude-3.7-Sonnet model with a heat map and line chart, maintained perfect consistency across all questions and temperature values, indicating a strong resistance to temperature-induced variability. These results highlight Claude's robustness in binary decision tasks. The line charts of LLaMA-v3.1-8B and ChatGPT-4o-mini, which can be seen in Figure 6(A) and 6(B), visualize the usual chaotic behavior we saw in other categories as well, although in this category, it can be seen as if the accuracy jumps up and down when looking at the line chart.

Logical Questions

The Logical Questions category was designed to assess the reasoning and inference capabilities of the models under different temperature settings.

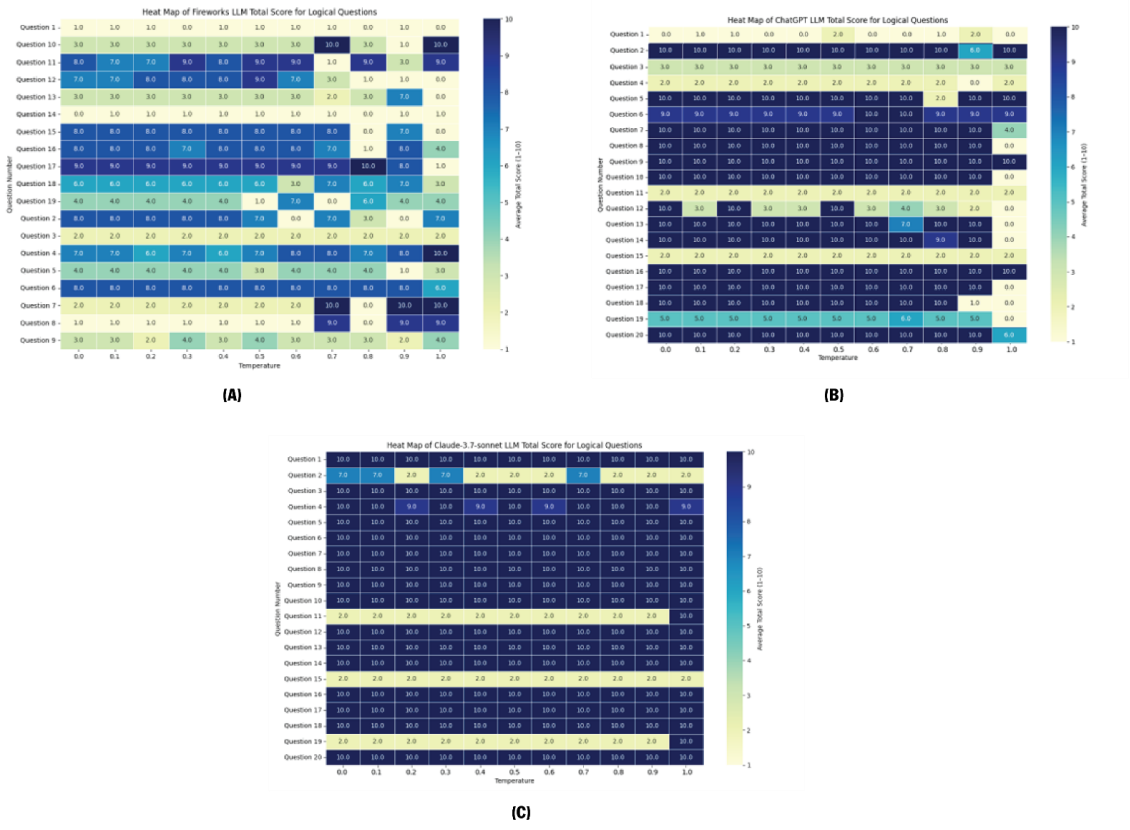

Figure 7. (A) Llama-v3.1-8b Heat map, **(B)** ChatGPT-4o-mini Heat map, and **(C)** Claude-3.7-sonnet Heat map

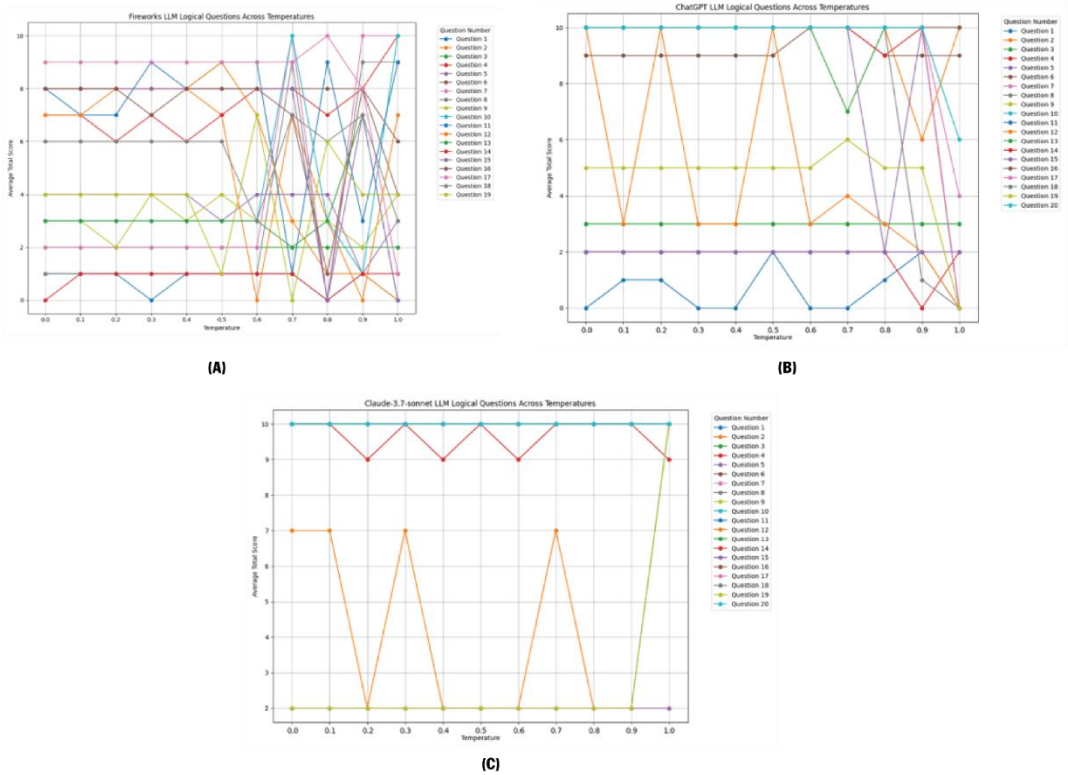

Figure 8. (A) Llama-v3.1-8b Line Chart, **(B)** ChatGPT-4o-mini Line Chart, and **(C)** Claude-3.7-sonnet Line Chart

With logical questions, LLaMA-v3.1-8B and ChatGPT-4o-mini models which are shown in Figures 7(A) and 7(B) proved to be highly consistent at low temperatures. Claude-3.7-Sonnet, shown in Figure 7(C) recorded an improvement at high temperatures, as compared to the others in Figures 7(A) and 7(B), which had a steady decline.

Both LLaMA and ChatGPT were highly unpredictable when the temperature was raised, and their consistency declined drastically which is seen in the line chart figures of LLaMA-v3.1-8B and ChatGPT-4o-mini shown in Figure 8(A) and Figure 8(B). The LLaMA-v3.1-8B model in Figure 8(A), appeared to behave more chaotically in higher temperatures which can be seen in the line graph. There was a significant difference in the quality of their answers to multiple queries. The performance of Claude shown in Figure 8(C), though generally strong, showed uneven variations, which did not seem to be affected by the temperature, and which occasionally worsened and other times improved in unpredictable manners.

General Knowledge

The General Knowledge category was included to evaluate the model's ability to recall and apply factual information across a broad range of topics. These questions primarily tested knowledge retrieval rather than complex reasoning or mathematical computation, providing insight into how temperature influences factual accuracy and response consistency. Heat maps were used to visualize the correctness of individual responses across different temperature settings, allowing for a detailed comparison of model performance.

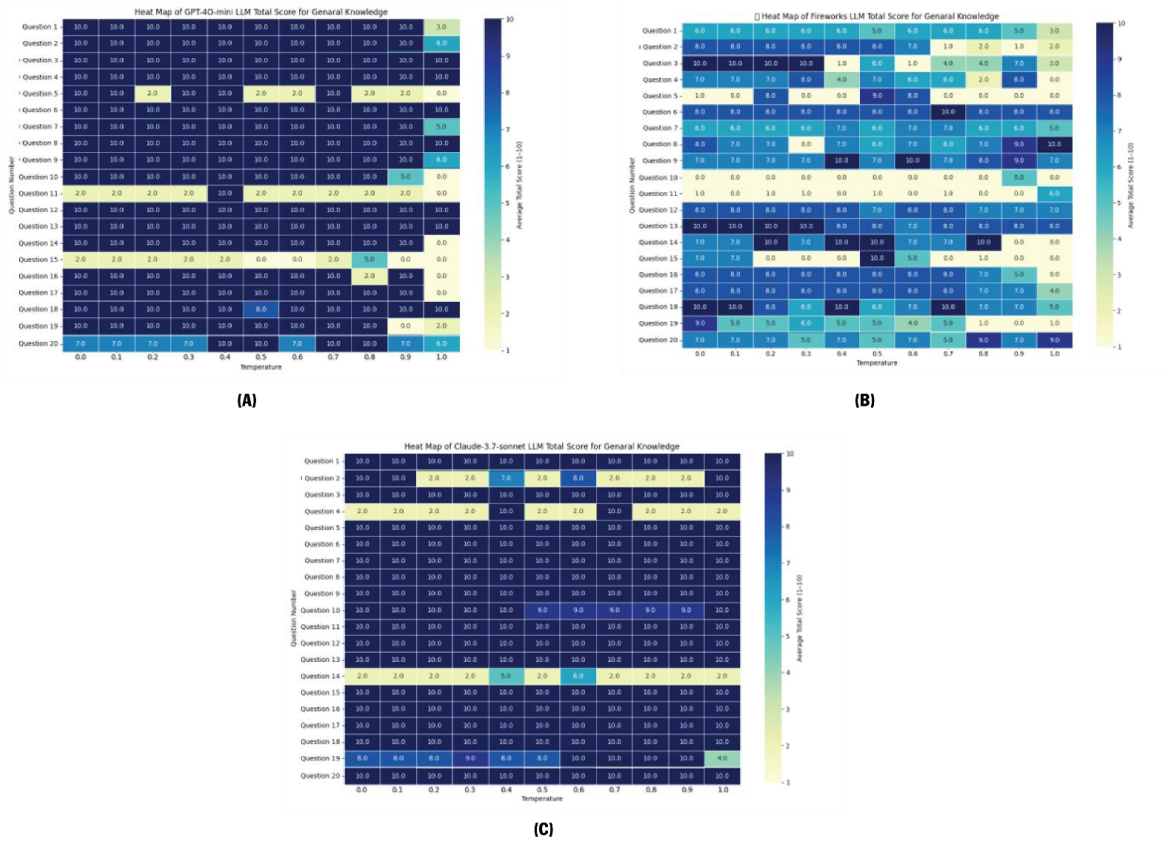

Figure 9. (A) Llama-v3.1-8b Heat map, **(B)** ChatGPT-4o-mini Heat map, and **(C)** Claude-3.7-sonnet Heat map

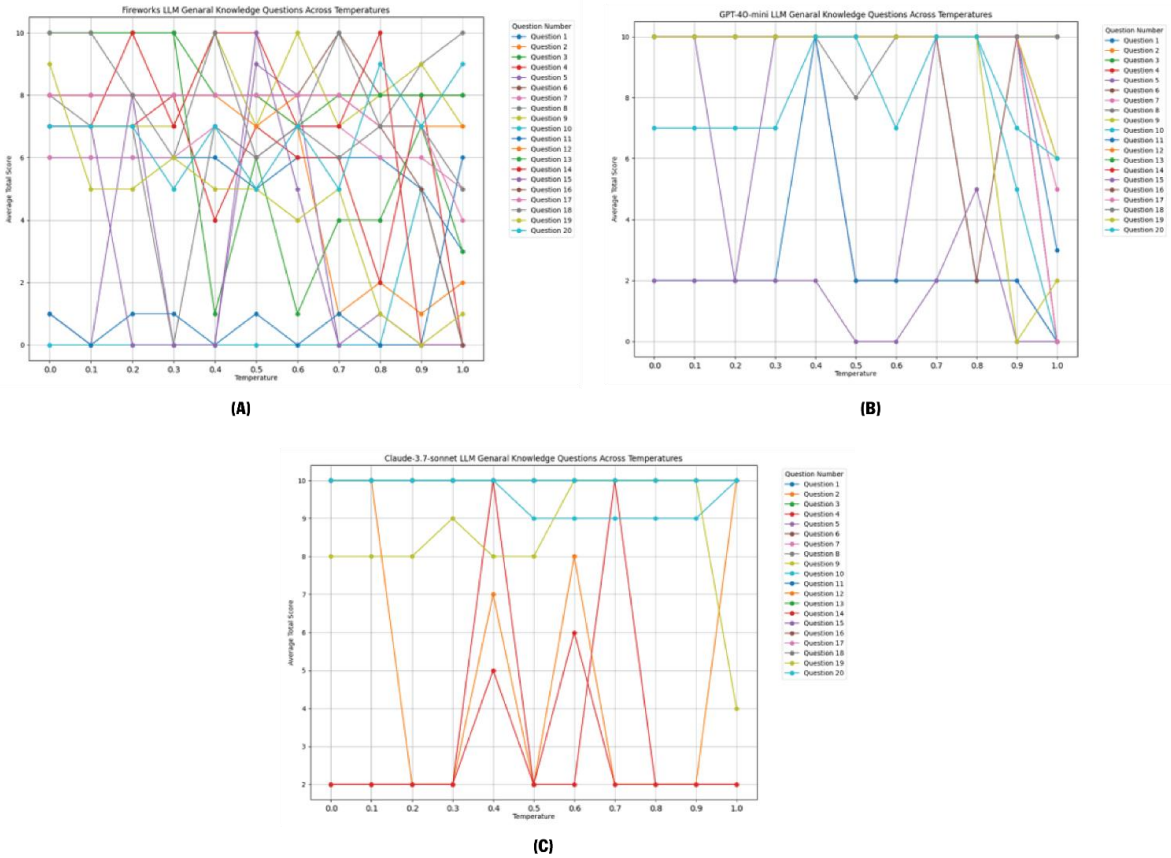

Figure 10. (A) Llama-v3.1-8b Line Chart, **(B)** ChatGPT-4o-mini Line Chart, and **(C)** Claude-3.7-sonnet Line Chart

As illustrated, the three models were very reliable on general knowledge questions at lower temperatures. LLaMA-v3.1-8B, Figure 9(A) and ChatGPT-4o-mini, Figure 9(B), were quite stable in those temperatures. But the Claude-3.7-Sonnet model outperformed the two, but showed some variation at elevated temperature conditions as can be seen by Figure 9(C). However, the scale of these changes was not as high as the other two models.

At higher temperature levels, LLaMA and ChatGPT displayed clear inconsistencies which are visualized by the line charts shown in Figure 10(A) and Figure 10(B), with accuracy dropping sharply or spiking unpredictably across several questions. The LLaMA model shown in Figure 9(C) shows chaotic behavior even from temperatures as small as 0.2. Claude, while generally more robust, showed irregular trends at mid-range temperatures, sometimes improving, other times degrading depending on the specific task as shown in Figure 10(C).

Overall Analysis

A broader analysis of the performance of the three large language models across the five question categories was conducted to identify overarching trends and behavioral patterns associated with variations in temperature.

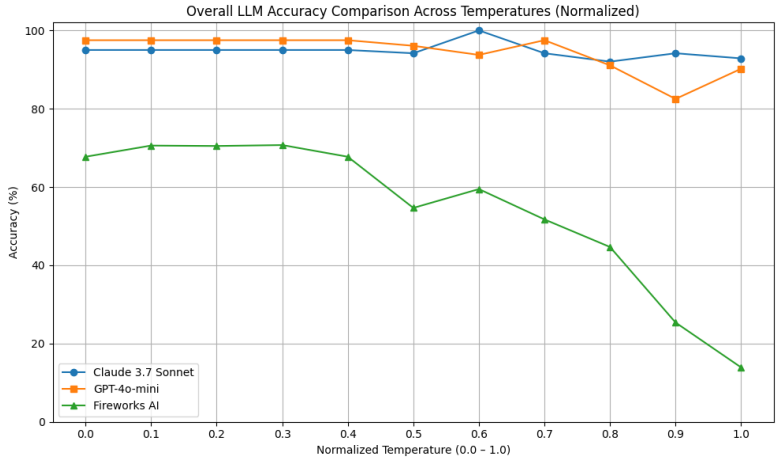

Figure 11. Overall Accuracy Mathematical Numerical

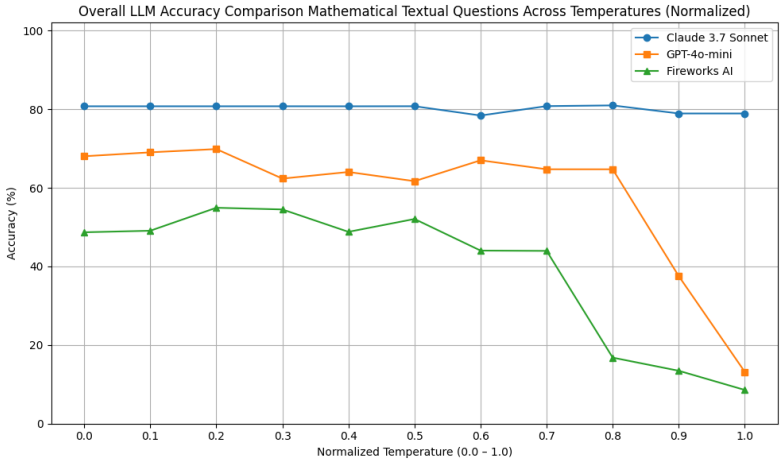

Figure 12. Overall Accuracy Mathematical Text-Based

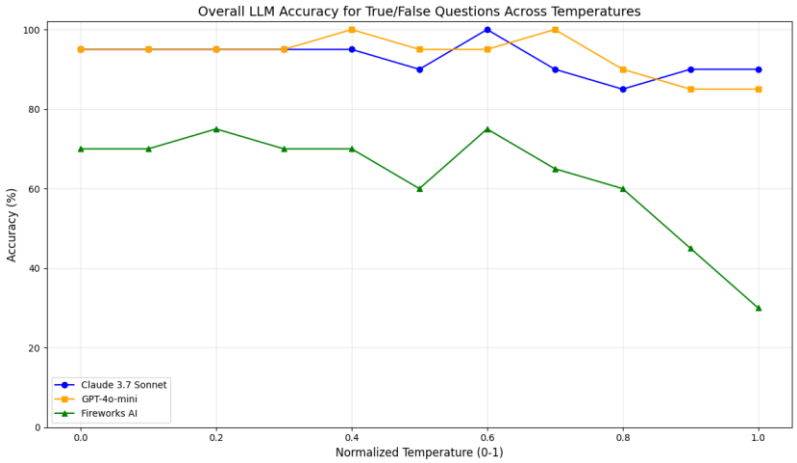

Figure 13. Overall Accuracy True or False

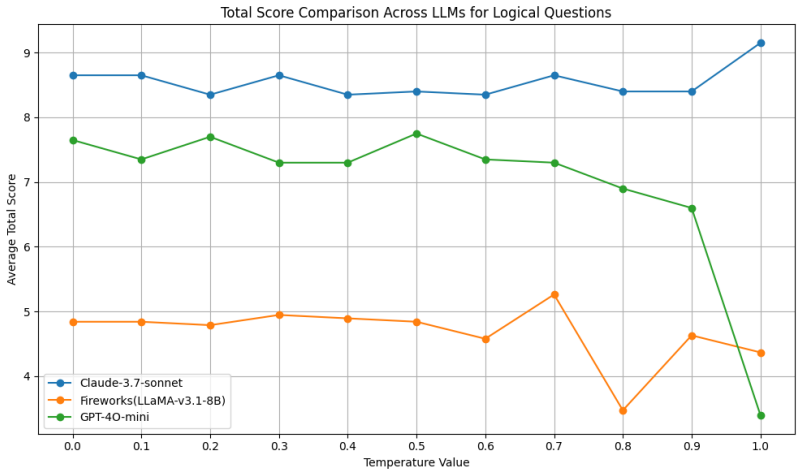

Figure 14. Overall Accuracy Logical Questions

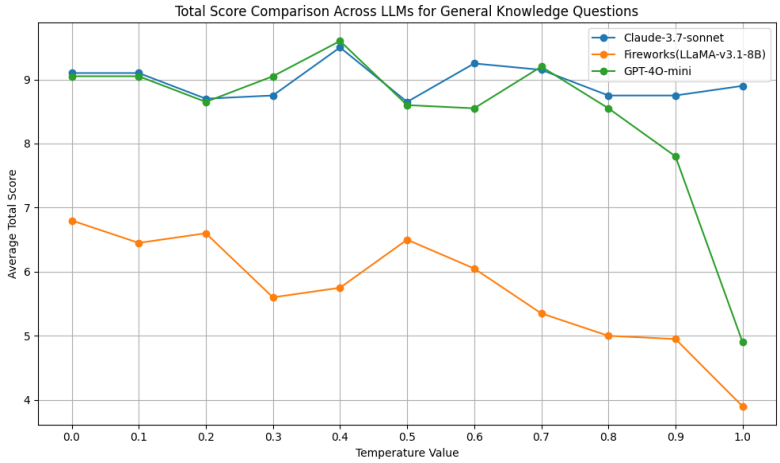

Figure 15. Overall Accuracy General Knowledge

The evaluation of model behavior across all five question categories reveals clear, category-dependent emergent-like patterns that are closely tied to temperature variation. While scale plays a role in overall model capability, our findings suggest that temperature can be a crucial and underappreciated driver of behavioral change in LLMs, triggering both constructive and destructive emergent-like traits depending on the model and task type.

Claude 3.7-Sonnet and GPT-4o-mini consistently exhibit high consistency across most temperature settings, particularly in well-structured tasks such as mathematical numerical, true/false, and general knowledge categories. In contrast, LLaMA-v3.1-8B, representative of a smaller-scale model, demonstrates significant volatility as temperature increases. This sensitivity manifests not as a smooth decline, but as sharp, nonlinear shifts in output quality, hallmarks of emergent-like instability.

Claude and GPT4o-mini have an almost perfect consistency until around 0.7 temperature, in the Mathematical Numerical category which is visualized by Figure 11. And above this point, GPT4o-mini can be seen to show a subtle decrease in performance, but LLaMA can be seen to suffer a significant decline at around a temperature of 0.5.

The Text-Based Mathematical category exerts certain requirements on semantic reasoning and explains the resulting differences as shown in Figure 12. The Claude model is mostly stable, and the GPT-4o-mini model becomes more and more unpredictable as temperature settings transition to the mid-range. The rate of semantic misinterpretations and reasoning hallucinations increase dramatically at these temperatures. LLaMA once again exhibits a dramatic collapse beyond temperature 0.8.

In the case of true/false items shown in Figure 13, models Claude and GPT-4o-mini show steady results. The LLaMA model is highly consistent at lower temperatures but becomes erratic beyond 0.6 temperature.

The Logical Reasoning category shown in Figure 14 highlights a unique emergent-like behavior in Claude. In contrast to the other two models, it becomes more accurate at times with an increase in temperature. This observation suggests that a controlled degree of randomness may actually enhance its ability to reason through abstract problems. ChatGPT-4o-mini and LLaMA are sensitive to the traditional deterioration with temperature increases; specifically, LLaMA has an increase in reasoning errors with higher temperatures.

In the general knowledge category which is visualized in Figure 15, both Claude and GPT4o-mini demonstrate high baseline performance, maintaining high consistency up to around 0.8 of temperature. Beyond this point, there is a performance drift on both models, but in Claude, variations are relatively small. LLaMA first shows decent reliability; however, its behavior capability decreases sharply with the increasing temperature, which underlines its limited applicability to the informational tasks that are not focused on a specific topic and are open-ended.

These results demonstrate that temperature alone can have an impact on behavior of LLMs, which expose emergent-like latent abilities and flaws in these systems. Claude 3.7-Sonnet stands out as the most resilient of all the models that were studied, and in this case, its ability to withstand a variety of prompts tied to a variety of problem categories and changes in stochastic sampling could be explained by its internal mechanisms. GPT-4o-mini can balance strength with flexibility in the generation of responses, but it shows detectable behavioral shifts in some semantic tasks when exposed to high temperatures. On the other hand, LLaMA-v3.1-8B demonstrates somewhat good performance in low-temperature settings but displays fragility in cases of higher temperatures, resulting in abrupt breakdowns and a decline in functionality.

This research aimed to examine how temperature as a stochastic sampling parameter affects the behavioral dynamics of LLMs. Opposed to the common beliefs that emergent behavior is tied to a model's scale, we have empirical results that suggest temperature alone could give rise to non-linear behavioral changes in even relatively small model architectures. This finding significantly reframes the understanding of emergence in generative AI.

Across five diverse question categories and 11 temperature settings, all three models, Claude 3.7 Sonnet, GPT-4o-mini, and LLaMA-v3.1-8B, exhibited clear behavioral patterns tied directly to temperature. Claude 3.7 Sonnet demonstrated the highest degree of stability, across most categories and temperatures, with cases where increased temperature unexpectedly affected the reasoning capabilities of the model. GPT-4o-mini followed with similarly high accuracy but showed semantic instability at mid-to-high temperature levels in text-based and logical tasks. LLaMA-v3.1-8B, the smallest model tested, showed emergent fragility: accuracy dropped abruptly beyond temperature 0.5 in several categories, coupled with erratic and incoherent outputs. These changes were not gradual, but non-linear, indicating true emergent-like traits driven by temperature rather than scale.

The findings suggest that emergent-like behavior might not be exclusive to large models, but rather a result of interaction between a model's architecture, its training patterns, and stochastic generation parameters. This insight breaks from prior work that correlated emergence primarily with increasing model size, and instead points toward temperature as a behavioral trigger capable of revealing latent model dynamics.

Furthermore, this study highlights the task-dependent nature of temperature effects. Models performed more reliably on structured tasks (e.g., true/false, numerical questions) and showed greater volatility in semantically complex or open-ended tasks (e.g., logical reasoning, general knowledge). Notably, some models like Claude-3.7-sonnet improved at higher temperatures on reasoning tasks, suggesting that controlled randomness can occasionally enhance performance, underscoring the importance of adaptive temperature tuning in practical applications.

Conclusion

This paper examined how temperature influences emergent-like behavior in LLMs. The research explored Claude 3.7 Sonnet, GPT-4o-mini, and LLaMA-v3.1-8B in five types of question datasets and eleven temperature ranges. The results supported the hypothesis: temperature alone can trigger emergent-like behaviors, independent of model scale. Key findings include a temperature-dependent effect in the form of sudden changes in behavior, changes in reasoning patterns, and category-specific impairment. These effects are more severe in smaller architectures, as in the case of LLaMA model. The implication of these observations suggests that emergent-like behavior can be a consequence of sensitivity to stochastic sampling and may not just be because of model scale. In practice, the results emphasize the need for temperature-aware design in LLMs, the use of adaptive sampling methods along with stability diagnostics can help alleviate the occurrence of erratic results. Introduction of unified standards is likely to increase the levels of transparency in model assessment. Modifying the temperature parameter makes the softmax function choose words with less probabilities for its predictions. Which leads to the model behaving unpredictably even ignoring user prompts, hence, showing signs of emergent-like behavior. However, the exact mechanical/ theoretical reasons for emergent-like behavior with temperature changes are outside the scope of this research project, as a study like that requires much more resources than what was available to us. This can be a topic that future researchers can expand upon, future work should scale this analysis to inside mechanisms, more models, agentic AI models and fine-tuned, domain-specific variants. Analyzing multi-turn reasoning could reveal how temperature impacts sustained coherence in dialogues.

Moreover, these insights could power a web-based observability toolkit, enabling developers and regulators to stress-test models under variable generative conditions. Finally, a temperature auto-tuning algorithm, responsive to task type or uncertainty, could offer real-time control over generation quality, making LLMs more reliable in real-world use cases.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This research did not receive external funding from public, commercial, or not-for-profit agencies. A portion of the support for this work was provided by the University of Sri Jayewardenepura.

Supporting Information

The source code used for the evaluation, along with the questions, prompts and outputs can be found on this GitHub page: https://github.com/L4Y3R/B05-Research-G17

References

[1] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W., Emergent Abilities of Large Language Models, **2022**. 10.48550/arXiv.2206.07682.

[2] Schaeffer, R., Miranda, B., and Koyejo, S., Are Emergent Abilities of Large Language Models a Mirage?. **2023**. 10.48550/arXiv.2304.15004.

[3] Renze, M., The Effect of Sampling Temperature on Problem Solving in Large Language Models. **2024**, 7346-7356. 10.18653/v1/2024.findings-emnlp.432.

[4] Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A., Is Temperature the Creativity Parameter of Large Language Models?. **2024**. 10.48550/arXiv.2405.00492.

[5] Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y., The Curious Case of Neural Text Degeneration. **2020**. 10.48550/arXiv.1904.09751.

[6] Brown, T.B., Mann, B., and Ryder, N., Language Models Are Few-Shot Learners. **2020**. 10.48550/arXiv.2206.07682

[7] Trusilo, D., Autonomous AI Systems in Conflict: Emergent Behavior and Its Impact on Predictability and Reliability. Journal of Military Ethics, **2023**. 22*(1)*, 2-17. 10.1080/15027570.2023.2213985.

[8] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., and Xie, X., A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology, **2024**. 15*(3)*, 1-45. 10.1145/3641289.

[9] Li, L., Sleem, L., Gentile, N., Nichil, G., and State, R., Exploring the Impact of Temperature on Large Language Models:Hot or Cold?, **2025**. 10.48550/arXiv.2506.07295.