

A Comparative Accuracy Analysis of Llama-3.3-70B and Qwen-3-32B Using the Groq API

1st Anubhav Shakya
School of Computing
Dublin City University, Dublin, Ireland
anubhav.shakya2@mail.dcu.ie

2nd Khushboo Kumari
School of Computing
Dublin City University, Dublin, Ireland
khushboo.kumari3@mail.dcu.ie

Abstract:-

In today's world, large language models are now frequently used in day-to-day applications. The development and testing of artificial intelligence (AI) systems has recently outgrown its significance. The goal of this project is to compare the reasoning accuracy and response consistency of two large language models deployed on the Groq platform: Llama 3 (70B) and Qwen 3 (32B). The aim is to test how these models interpret the user's query and respond to the questions which are logical and require analytical thinking rather than simply memorizing the facts.

To accomplish this project, we are using a JavaScript based interface that is the "Ancient Brain" platform, which communicates with Groq's inference API to evaluate this project and give us the required outcomes. The system retrieves the answers from both models for the comparison after the user manually enters the questions. The study offers insights into how various model architectures handle structured reasoning tasks by evaluating each response's accuracy and clarity.

This project is to emphasize a practical grasp of the behaviour of AI models, evaluation techniques, and comparative analysis. The project's results show the advantages and disadvantages of existing open-source large language models and advance knowledge of their capacity for controlled reasoning.

Introduction:-

Artificial Intelligence is now widely used for translation tasks, reasoning, some problem-solving tasks and for content generation. These systems are highly capable of doing such types of tasks with high accuracy but it is not always accurate or consistent. The LLM model faces difficulties when the prompt that is given to it is ambiguous or has logical traps. Therefore, evaluation of LLM models has become an important area of study because these systems are now being used in real-world applications.

This project focuses on analysing the accuracy in the output of two LLMs – Llama-3.3-70B-Versatile and Qwen-3-32B which has been accessed through Groq API. The aim of this experiment is to test and compare how these two LLMS perform on different sets of logical and reasoning questions. Rather than focusing on the simple questions, we have used the dataset that has the mixed questions such as logical puzzles, counting questions, trick questions and some translation tasks. We choose these categories of questions because these are known to expose the strengths and weaknesses of the AI models.

By asking the same set of questions to both the models and analysing the outputs, the project aims to identify the areas where one model performs well and give a more accurate answer than the other one. This comparative evaluation does not only provide the insights in the behaviour of the models but also highlights the broader issue which is related to the AI reliability and trusting the AI.

Methodology:-

1. Dataset Construction:-

The dataset of custom questions was created to test the range of cognitive and linguistic abilities of the models. The dataset consisted of the logical puzzles, counting questions, pattern-recognition, trick questions and some translation type questions. We intentionally selected these categories of questions because they reveal the different types of weaknesses in the LLMs. For example, we used questions such as "*How many r's in Strawberry*" to test the counting accuracy and some questions like "*Translate to Spanish: I am excited*" to test the semantic understanding in translation tasks rather than just translating word to word. For each question, there are answers that are more accurate. These answers were the ground-truth answers to test the output of the models.

2. System Implementation:-

To conduct this experiment of evaluation, we used a JavaScript based world that was developed on the Ancient Brain platform. This world interacts with the Groq API and sends each question to both the models using their API:

- **Llama-3.3-70B-Versatile**, and
- **Qwen-3-32B**

The world shows the output of both the models side-by-side for comparison.

A key component in this system involves a step in processing the model's output. The Qwen model was giving the output along with the internal reasoning traces enclosed in <think>...</think> tags. These tags and the reasoning were removed programmatically before the evaluation of these models and only the output of both models was displayed to the end user.

3. Evaluation Criteria:-

Since the dataset contains multiple types of questions, the different evaluation rules were used depending on the type of questions asked.

Objective Logic / Counting Questions: For questions such as “*How many months have 28 days?*” or “*What is the next number in the sequence?*”, the response was marked correct if the model generated the output with the exact number in its answer.

Translation Task: For the translation tasks, we used google translate to check the meaning of the translated phrase by these models. For example, the correct translation of “*I am excited*” to Spanish is “*Estoy emocionado/a*”. So, if the model generated the output that contains these phrases, then it was marked as correct.

Trick Questions: There were some trick questions asked to these models specially to test the cognitive skills of these models because the models fall for these types of trick questions. For example, questions like “*What begins with T, ends with T, and has T inside it?*” and “*If there are 12 fish and 5 drown, how many are left?*” are more of the logical questions. These questions were scored based on the understanding of these trick logical questions.

However, as these questions do not have any right or wrong answers, the main purpose of the evaluation was not to assign the right or wrong tags to the models but to observe which model generated the more accurate answer than the other. In such a case, both the models may generate the answer correctly but the way, style, clarity and level of reasoning in the response shows the difference in the response of each model.

Results:-

A total of eight questions were asked to these models, covering the logic, counting, translation, and reasoning to evaluate the accuracy of the Llama-3.3-70B-Versatile and the Qwen-3-32B models. We asked the same set of questions to both the models and then we compared the response of these models against the expected answer. We measured the accuracy of the models based on the logic they used to generate the answers because the answers they were generating were correct. Table 1 is the summary of the responses of each model.

Question Type	Total Questions	Llama Correct	Qwen Correct
<i>Translation</i>	2	2	2
<i>Logic / Counting</i>	4	2	4
<i>Riddles / Trick Questions</i>	2	2	2
Overall	8	6	8

The results show that the Qwen model achieved higher accuracy in generating the more correct answers than the Llama model. The Qwen answered all 8 questions correctly while the Llama only answered 6 correctly.

Although both models often produced the correct responses, their reasoning style, presentation format and the depth of the explanation were different in all the questions. The differences were not in the accuracy but it was in how each model is approaching a certain problem. We will see the qualitative difference in the responses of the models.

1. Translation Tasks Example: “Translate to Spanish: I am excited.”

- Llama: It provided the translation for both, male and female (“Estoy emocionado / emocionada”), showing that it has great linguistic awareness.
- Qwen: It only generated the one option (“Estoy emocionado”), that means it provides a more simple response.

Llama gives the richer and complete translations with the gender variants also but the Qwen only produced the short and direct response.

2. Counting and Logic Questions Example: “How many r’s in Strawberry?”

- Llama: It gave the incorrect response that there are only 2 r’s in the word ‘Strawberry’ and it gave a very confusing explanation about it.
- Qwen: It spelled the word out letter by letter and then counted the frequency of the letter ‘r’ in the word and then gave the response 3.

The screenshot shows a web browser window with the URL [run.ancientbrain.com/run.php?world=Groq API](https://run.ancientbrain.com/run.php?world=Groq%20API). The page title is "Compare Two AI Models on Logical Reasoning". It says "This world compares llama-3.3-70b-versatile and qwen3-32b using the Groq API." There are two steps: Step 1: Enter Groq API Key (with a success message "API key set successfully!") and Step 2: Enter Your Question (with the question "How many r's are in the word 'strawberry?'"). The Qwen3 32B section shows step-by-step reasoning: "The word 'Strawberry' is spelled S-T-R-A-W-B-E-R-R-Y. Breaking down each letter: 1. S, 2. T, 3. R, 4. A, 5. W, 6. B, 7. E, 8. R, 9. R, 10. Y. The letter 'R' appears at positions 3, 8, and 9. Therefore, there are 3 instances of the letter 'R' in 'Strawberry.'". The Llama 3.3 70B Versatile section provides a summary answer: "Answer: There are 3 R's in the word 'Strawberry.'".

Qwen used the step-by-step reasoning in generating the answer while the Llama attempted to summarise the answer without counting as shown in the above image.

Discussion:-

Across the questions, it was noted that the Llama is generally producing the answers in more fluently and natural sounding language. However, sometimes it shows the signs of lack of reasoning and minor hallucinations in some tasks where the explanations become repetitive tasks. On the other hand, the Qwen generates a more structured and step-by-step explanation for each question asked to it. The response of Qwen was more logical and the reasoning was typically more stable than Llama. Despite both the models generating most of the answers correctly, the reasoning style and the clarity of explanation were different in both the models.

In addition to these differences, the two models also demonstrated the difference in how much detail they are giving in their responses. Qwen provides longer answers with the big explanation and justifications of each answer even if the answer is straightforward. It also gives the internal thinking of the model in the response. It shows the transparency in reasoning in generating the answers. On the other hand, Llama generates the concise output with less explanations and more focus on the result.

Conclusion:-

The comparative evaluation of Llama-3.3-70B-Versatile and Qwen-3-32B highlights that both the models are capable of generating the accurate responses for the logical and reasoning types of question but they have their strength in different domains. Llama is great in the tasks that require natural language fluency and generates concise answers which makes it more effective. While the Qwen model excels in performing on the logical and reasoning questions and also gives the proper and structured reasoning of the answers. This behaviour of this Qwen model makes it more trustworthy. These findings from this project indicate that the selection of the models should be based on the use case and on the nature of the question to ask. Ultimately, this project shows that the evaluation of these AI models should not only be based on the correctness of the response but also on the logical understanding of these models.

References:-

Groq Inc. (2024) *GroqCloud API Documentation*. Available at: <https://console.groq.com/docs>

Meta AI (2024) *Llama 3 Model Card and Technical Report*. Available at: <https://ai.meta.com/llama>

Alibaba Cloud (2024) *Qwen3 Model Card*. Available at: <https://github.com/QwenLM/Qwen3>

Humphrys, M. *Ancient Brain — JavaScript AI Worlds*. Available at: <https://ancientbrain.com>

Creswell, A., Shanahan, M. and Higgins, I. (2022) *Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning*. arXiv preprint. Available at: <https://arxiv.org/abs/2205.09712>