



**Analyse Large language models and evaluate their usage inside CORE-QA. Which model performs better? Which model is the best compromise in terms of cost, accuracy, and performance?**

Obiogbolu Kaobimdi Ian  
09/24

## Abstract

Individual researchers spend hours in libraries, reading books, and looking through various articles to find the perfect information. It can take a while, and you might still end up with confusing answers. Technologies such as CORE, which is seen as a research dictionary containing millions of research articles in different kinds of fields, like science, business, and arts, serve as a one-stop solution for finding verified answers to research questions. However, searching this vast knowledge source to find the right answer can be challenging. By combining CORE with a suitable Large Language Model (LLM), we can provide well-detailed information and credible references to find the best answers.

This report explores how CORE and different LLMs handle research questions in various disciplines and compares LLMs to how individuals do research. The objective is to evaluate the response from the existing models in terms of comprehensiveness, trustworthiness and usefulness and to analyse if the LLMs understand the questions being asked. Based on the comparison results, we identify the best and the least performing LLMs for question answering for CORE.

# Table of Contents

Abstract .....	1
Introduction .....	3
Related Works .....	4
Methodology .....	4
Pricing.....	6
Evaluation .....	6
Challenges .....	8
Result .....	9
How do we evaluate different large language models? .....	10
Testing other various large language models .....	10
Conclusion.....	11
Future Work .....	11
Appendix.....	12

# Introduction

The combination of CORE with LLMs represents a significant shift in research methodologies, offering a blend of speed, adaptability, and comprehensive access to information. We have centred this research on 2 main LLMs which are:

- anthropic.claude-3-sonnet-20240229-v1:0
- anthropic.claude-3-haiku-20240307-v1:0

These models have been tested repeatedly along with some other models. We ran this test to analyse which LLM would be the most suitable integration with CORE. CORE plays an important role in providing the top 5 citations over 237M open-access articles collected from 11K data providers around the world. Out of numerous disciplines in the world, we selected the following subjects which would elaborate this process easily,

**(1) Biology** describes everything that goes on in our daily lives, including how our bodies work and how we interact with the natural environment. We chose biology because we want to study the natural wonders that inspire scientific research and advancement in domains like environmental conservation, agriculture, and medicine.

The goal of studying **(2) Information technology** is to comprehend the rapid growth of technology and how it affects our personal and professional lives. It's amazing how technology changes how we work, play, and learn while bridging the gap between science and ordinary life.

By diving into **(3) Literature**, we are interested in how storytelling shapes our perceptions of reality and challenges us to see the world from different perspectives. I am not really a fan of pure literature but had to put humanities into consideration and examine. Finally, we investigate the field of **(4) Social sciences**, to uncover the forces that shape our societies and culture.

In summary, the above disciplines hold the keys to understanding real-life challenges and how we adapt. We have chosen to focus on **biology, information technology, literature, and social science** because These fields also capture the essence of what it means to have a unique blend of science and humanities.

In this evolving landscape, CORE and LLMs will partner in the research process, creating a more customised technique for getting research answers. As we move forward, the question isn't just about which method is superior but how they can complement each other to create more efficient and insightful research.

## Related Works

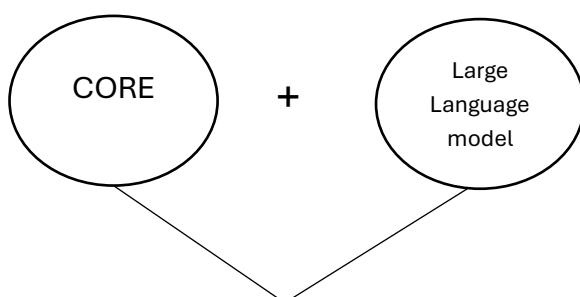
Considering the data source of this analysis, firstly we have come to understand how they all work. CORE (<https://core.ac.uk/>) currently contains 306M open-access articles collected from 11K data providers around the world. CORE aims to gather an aggregate of all open-access content distributed across different systems, such as repositories and open-access journals, enrich this content using text mining and data mining, and provide free access to it through a set of services.

On the other hand, LLM is a type of [intelligence](#) algorithm that uses [learning](#) techniques and massively large data sets to understand, summarise, generate and predict new content. Nonetheless, the future of LLMs will likely remain bright as the technology continues to evolve in ways that help improve human productivity.

CORE can identify relevant topics and keywords, while LLM can generate high-quality content aligned with those insights.

## Methodology

The main aim of combining both CORE and a specific model is to be able to provide quality answers that have genuine citations attached to them. For the citations, they are research works from real authors, part of academic publishers, institutional repositories, open-access journals, and subject repositories. CORE provides open-access research papers including full text and metadata.



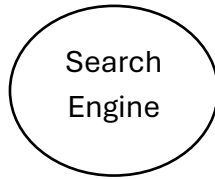


Fig 1: The background view of combining CORE and LLM.

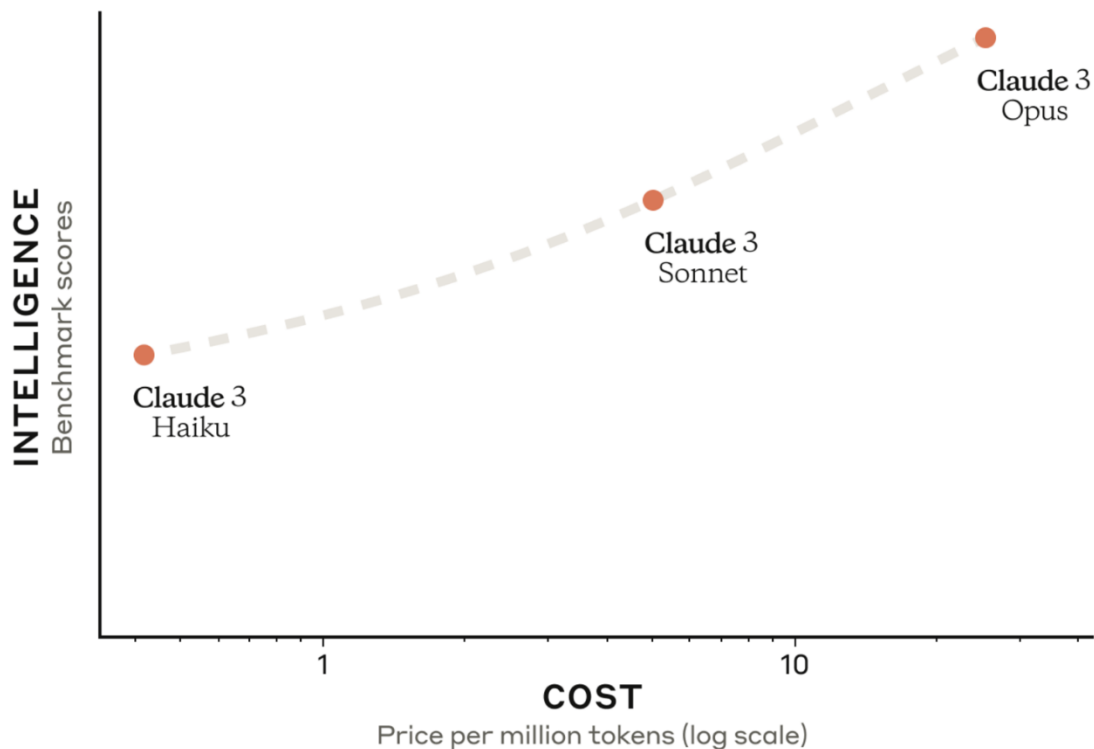
This brings us down to the main question, which is

**What model is best to use? And why?**

There are various models, and they may differ based on their attributes and how they work. We are looking forward to combining CORE which itself is open access to a model that could provide a well comprehensive and detailed answer without any fees or at least the minimum. We have done numerous tests with various LLM and practically drawn 2 main models.

- anthropic.claude-3-sonnet-20240229-v1:0
- anthropic.claude-3-haiku-20240307-v1:0

Claude 3.5 Sonnet sets new industry benchmarks for graduate-level reasoning (GPQA), undergraduate-level knowledge (MMLU), and coding proficiency (HumanEval). It shows marked improvement in grasping nuance, humour, and complex instructions and is exceptional at writing high-quality content with a natural, relatable tone. Claude 3.5 Sonnet operates at twice the speed of Claude 3 Opus. This performance boost, combined with cost-effective pricing, makes Claude 3.5 Sonnet ideal for complex tasks such as context-sensitive customer support and orchestrating multi-step workflows.



Source: <https://www.anthropic.com/news/claude-3-family>

Fig 2: The diagram describes the Claude 3 model family.

## Pricing

The Claude family includes three models: Claude 3 Haiku, the fastest and most compact model for near-instant responsiveness; Claude 3 Sonnet, the ideal balanced model between skills and speed; and Claude 3 Opus, the most intelligent offering for top-level performance on highly complex tasks.

Anthropic models	Price per 1,000 input tokens	Price per 1,000 output tokens
Claude 3.5 Sonnet**	\$0.003	\$0.015
Claude 3 Opus*	\$0.015	\$0.075
Claude 3 Haiku	\$0.00025	\$0.00125
Claude 3 Sonnet	\$0.003	\$0.015

Fig 3: The table above explains the pricing system of the anthropic models per input and outputs.

Considering the price list, it appears to be fair enough. The pricing seems to take into account various factors that typically influence costs, such as quality, and value provided.

We use PyCharm to write the code for CORE + LLM. The tool provides a strong and adaptable research solution for a wide range of fields. This dynamic and effective research tool combines the greatest features of both technologies by personalising keyword selection followed by answer generation.

Core API ensures access to extracting academic papers and keywords. LLMs rely on API calls to interact with the model and provide an elaborate answer. By using Python, CORE, and LLM we're not just conducting research; we're transforming the way we think about information and how we can creatively solve problems.

## Evaluation

We selected 5 questions from each discipline and ran a test to compare which model could possibly bring a perfect result to the outcome. Following are the three metrics which we use for evaluating the model.

- Comprehensiveness: How comprehensively is the question answered?
- Trust: How trustworthy is the answer?
- Utility: How useful is the answer?

We started by gathering specific research questions from various disciplines and then setting a prompt with CORE and for each model we generated a response for every search. This is followed by the evaluation of the quality, reliability, and trustworthiness of the information obtained. By analysing and comparing the responses, we assess the

strengths and weaknesses to compare each outcome, followed by summarising the results.

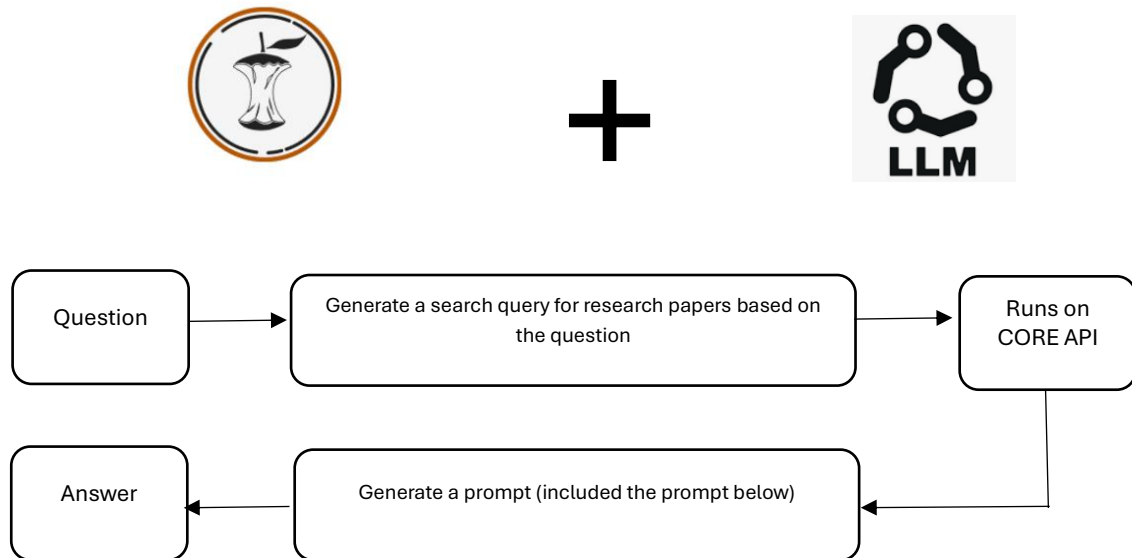


Fig 4: CORE + LLM workflow

Generating the search query based on the papers is more likely the job of the large language model. It sends a message to CORE API ensuring it displays this role to the next act.

```
def generate_search_query(input_request):  
    messages = [  
        {"role": "user", "content": "Generate a search engine query for a research paper based on the question. "  
        "Prioritise the most important keywords and add synonyms to focus the search. "  
        "Ensure that the response contain only the query and no other extra text"  
        "The answer should be no longer than "  
        "80 words."},  
        {"role": "assistant", "content": f"{input_request}"},  
    ]
```

In addition to that, it generates a prompt of what the result will entail as seen below

```
def generate_answer(input_request, search_results):  
    global final_answer  
    messages = [  
        {"role": "user", "content": "Generate a comprehensive answer (but no more than 160 words) "  
        "for a given question solely based on the provided search results in the format "  
        "{url:$url, abstract:$abstract}. "  
        "You must only use information from the provided search results."  
        "Use an unbiased and journalistic tone. Combine search results together "  
        "into a coherent answer. "  
        "Do not repeat text. Cite search results using the url provided and the [N] "  
        "notation. Only "  
        "cite the most relevant result that answer the question "  
        "accurately. If different results refer to different entities with the same "  
        "name, write separate answers for each entity."},  
        {"role": "assistant", "content": f"{input_request}"},  
        {"role": "user", "content": f"{json.dumps(search_results)}"}  
    ]  
    print(json.dumps(messages))  
    response = llm.invoke(messages) # return a response to the prompt  
    response_body = response.content
```

In summary, both models have unique strengths, Claude's Sonnet and Claude's Haiku. Claude Sonnet excelled in providing comprehensive answers and adhered closely to the prompt instructions, offering a robust conversion of queries. Claude Haiku is also effective but less proficient in query conversion, resulting in fewer responses. The differences highlight Claude Sonnet's ability to deliver well-



comprehensive writing and provide an answer while Claude Haiku might require further refinement to match that level of performance, which was a challenge to completing the test.

## Challenges

One major challenge we encountered while running the test was converting the query using Claude Haiku. While it provides valuable information, it often falls short of delivering a response. We discovered two solutions regarding that,

- Trying to customise the prompt to be more comprehensive
- Try to fix the code better so that the API understands the prompt

By changing the prompt, we then asked some problematic questions, and it still came out with “not enough results”. We realised it wasn’t coming from the prompt but could be the code. We went further to customise the prompt in a forceful way to avoid error while providing the answer, but it wasn’t just the only problem.

Specifically, then adjusted the structure of the API calls to ensure the query was being communicated more effectively. This involved:

- **Refining the Input Format:** we restructured the data we sent to the API, ensuring that the query was presented in a clearer, more logical sequence. This helped the model grasp the intent of the prompt better.
- **Improving Error Handling:** we introduced additional error-checking mechanisms within the code to catch and correct any misinterpretations before the API processed the query.

By making these changes, we were able to significantly improve Claude Haiku’s ability to understand and respond to queries, leading to more reliable outcomes. This experience also highlighted the importance of precise communication between the application and the API, which is crucial for maximising the performance of any language model.

After improving the error of the Claude haiku, to differentiate this improved version from the original, we renamed it "Claude Haiku Updated." This helped prevent any confusion between the original and the revised model, ensuring that we could clearly track improvements and record any other changes if they may apply.

## Result

The benchmark results originate from the test run across the 4 disciplines listed above, red interprets the least favourite whilst amber is good, and green is excellent. I have fixated this test on the comprehensiveness and detailing of the answer to come up with an overall review.

Answers ranked on

- Comprehensiveness
- Trustworthiness
- Utility

### Haiku

Domain	Comprehensive	Trust	Utility	Mean
Biology	7.5	9	8.5	8.3
Info Tech	9.5	9.5	9	9.3
Literature	-	-	-	-
Social Science	-	-	-	-

**=4.4**

*Table 1: The result shown is using Claude's haiku (Note: Where it says null, it was recorded as "not enough response", as this was one of the challenges faced but was updated and renamed in **Table 3**)*

### Sonnet

Domain	Comprehensive	Trust	Utility	Mean
Biology	8	9	9.5	8.8
Info Tech	9.5	9.5	9.9	9.6
Literature	8	7.5	8.5	8.0
Social Science	9.5	9.9	9	9.4

**=8.95**

*Table 2: The result shown is using Claude's sonnet.*

### Claude's Haiku Updated

Domain	Comprehensive	Trust	Utility	Mean
Biology	7.5	9	9	8.5
Info Tech	9.5	9.5	9.5	9.5
Literature	7.5	7.5	8.5	7.8
Social Science	9.5	9.9	9.5	9.6

**=8.85**

Table 3: The result shown is using Claude's haiku updated (Note: This version of Claude's haiku is the revised version).

### Overall Benchmark













	Biology	Information Tech	Literature	Social Science
Claude Haiku	8.3 	9.3 	0.0 	0.0 
Claude Sonnet	8.8 	9.6 	8.0 	9.4 
Claude Haiku Updated	8.5 	9.5 	7.8 	9.6 

Table 4: Overall benchmark of the different models, judging the trustworthiness of the outcome, the usefulness of the outcome, and the citations.

## How do we evaluate different large language models?

In evaluating LLM, there are numerous ways to do that starting from an **individual** approach. This simply is an individual comparison to different types of LLM, it could be any specific criteria. **Popularity** aside, performance is a key approach to this. In the sense that we need to measure and compare the **performance** it provides and evaluate the accuracy and efficiency of models in answering questions. For instance(example)

Another primary way of evaluating LLM's is by logical **reasoning**. It may involve testing the model's execution on tasks that require understanding and using logical sequences. This can also be extended to deductive and inductive reasoning. For instance(example)

**Accuracy** evaluates how closely the outputs of LLMs match verified real-world facts; this is important as must assess the models' capacity to deliver consistent, truthful information free from mistake or error. By doing this we need to identify instances in which LLMs generate information that is either factually wrong or not based on reality. For instance(example)

### Continuous learning and peer communication

Learning from any assessment you must have used to test the model and discussing among your peers if there is any contribution to evaluating LLMs.

## Testing other various large language models

Testing various large language models to evaluate what could be considered the best for evaluation. The benchmark results mainly centred on a logical question, a typo error

question and how trustworthy the output is. The colour red is interpreted as bad, whilst yellow is good, whilst amber is very good, whilst green is excellent, and green is very excellent



	Testing based on typo errors & Response	Testing based on accuracy & truthful information	Testing based on performance (speed, summary)
Gemini	Yellow	Light Green	Light Green
Open Ai	Light Green	Light Green	Light Green
Claude Ai	Light Green	Yellow	Yellow
Nova Ai	Yellow	Red	Red
Deep Ai	Red	Yellow	Yellow

Table 5: This result shown is from an earlier test using some other large language models for analysis. (The REST of the analysis is attached in the Appendix below)

## Conclusion

After analysing various large language models within the CORE-QA framework, it is clear that Claude Sonnet stands out as the best option. This model excels in compatibility with our specific needs, offering more precise and careful query interpretation compared to other models. Claude Sonnet’s ability to deliver accurate and comprehensive responses makes it a reliable choice for complex tasks. Additionally, its pricing is reasonable, reflecting its affordability and compatibility, making it a cost-effective solution.

In contrast, Claude Haiku posed significant challenges, requiring frequent prompt adjustments and response cleanups, which ultimately affected its efficiency. Through benchmarking and detailed comparison, it became evident that Claude Sonnet offers the best balance between cost, accuracy, and performance. Therefore, Claude Sonnet is the most suitable model for our CORE-QA application, ensuring high-quality results with minimal complications.

## Future Work

1. Introducing a feature that allows users to re-edit their questions after submitting them can significantly improve the accuracy and relevance of the responses. This would enable users to refine their queries based on initial results, leading to more targeted and useful outcomes

2. Increasing the word limit in the prompt backend can provide more context to the language model, allowing it to generate more comprehensive and nuanced responses. This would be particularly beneficial for complex queries that require detailed explanations.
3. Adding a "Copy All" button would allow users to quickly copy the entire response with one click, making it easier to transfer the information to other documents or applications. This feature will enhance user experience by simplifying the process of gathering and utilising the generated content.

We can greatly improve the system's usability, accuracy, and general efficiency by including the above features, which will eventually result in a more reliable and user-friendly CORE-QA platform.

## Appendix

Testing other various large language models, these questions have been asked and attached *are the responses*:

### Testing based on Typo/Logical

**Question: If Bryan was born in the year 2000, and his brother Ethan is 5 years younger than him. How old is Brayan**

#### Gemini:

You already established Bryan's age! Bryan, born in 2000, is still Bryan. We don't need to recalculate his age.

Therefore, Bryan is currently **24 years old**. (2024 - 2000 = 24)

#### OpenAI:

To determine Bryan's current age, we need to know the current year. As of 2024, Bryan, who was born in 2000, is:

2024-2000=24 years old  
 $2024 - 2000 = 24 \text{ \textit{years old}}$   
2024-2000=24 years old

So, Bryan is 24 years old.

#### Claude Ai:

There's a small error in your question - you mentioned "Bryan" at the beginning but "Brayan" at the end. I'll assume you meant Bryan throughout.

To solve this problem, let's break it down step by step:

1. Bryan was born in 2000.

2. The current year is 2024 (as per the date information provided to me).

3. To calculate Bryan's age:  $2024 - 2000 = 24$  years old

So, Bryan is 24 years old.

However, the question asks about Ethan's age. Ethan is 5 years younger than Bryan.

4. To calculate Ethan's age:  $24 - 5 = 19$  years old

Therefore, Ethan is 19 years old.

Would you like me to clarify or explain any part of this calculation?

#### **Nova Ai:**

If Bryan was born in the year 2000, he would be 24 years old now. Since his brother Ethan is 5 years younger than him, Ethan would be  $24 - 5 = 19$  years old.

#### **Deep Ai:**

Bryan was born in 2000. If Ethan is 5 years younger, then he was born in 2005.

#### **Testing based on Trustworthiness/Comprehensiveness**

**Question: what way does the world clock go and why?**

#### **Gemini:**

Clocks around the world, both physical and digital, traditionally go in a **clockwise** direction. This convention likely stems from the way sundials worked in the Northern Hemisphere.

Early timekeeping relied on sundials, which track the sun's movement across the sky. In the Northern Hemisphere, the sun's shadow rotates clockwise throughout the day. When mechanical clocks were invented, they were designed to mimic this motion, making the hands turn clockwise as well.

While the sundial connection is the leading theory, it's important to note that clocks could have conceivably gone the other way. The choice of clockwise might have simply become the standard due to its early adoption.

-

## **OpenAI:**

The "world clock," or the standard way that clocks display time, moves in a clockwise direction. This directionality has historical and practical origins:

### Historical Origins:

**Sundials:** The earliest clocks were sundials, which were used in the Northern Hemisphere. The shadow on a sundial moves in a clockwise direction because the Sun appears to move from east to west in the southern part of the sky. Early clockmakers designed mechanical clocks to mimic the motion of the shadow on a sundial, which led to the clockwise convention.

**Mechanical Clocks:** When mechanical clocks were developed in Europe, the hands were designed to move in the same direction as the shadows on sundials. This was a natural choice since sundials were the primary timekeeping devices at the time.

### Practical Considerations:

**Standardisation:** Once the convention of clockwise motion was established, it became standard practice worldwide. Changing this convention would cause confusion and require redesigning clocks and watches.

**Design and Functionality:** Clock mechanisms, which rely on gears and other mechanical parts, were designed to move in a specific direction. Reversing the direction would necessitate redesigning the internal mechanisms of clocks and watches, which is impractical.

### Cultural Influence:

**Global Adoption:** As mechanical clocks spread around the world, different cultures adopted the clockwise convention. Today, this convention is universally recognized and used.

### Summary:

The clockwise motion of clocks is a result of historical developments, specifically the way shadows moved on sundials in the Northern Hemisphere. This direction was

adopted by early clockmakers and has remained the standard due to practicality and widespread acceptance.

### **Claude Ai:**

The world clock, also known as the international time zone system, generally moves from west to east. This means that as you travel eastward around the globe, the time becomes later, and as you travel westward, it becomes earlier. Here's a brief explanation of why:

1. **Earth's rotation:** The Earth rotates on its axis from west to east, which means the sun appears to move across the sky from east to west.
2. **Sun's apparent movement:** As the Earth rotates, different parts of the planet face the sun at different times, creating day and night cycles.
3. **Time zone progression:** To maintain a consistent relationship between clock time and the sun's position, time zones are set up to progress eastward.
4. **International Date Line:** This imaginary line, roughly along the 180-degree meridian in the Pacific Ocean, is where the date changes. When you cross it going west, you add a day; going east, you subtract a day.

This system allows for a logical progression of time around the globe, with each time zone typically being one hour apart from its neighbours (though there are exceptions).

Would you like more information on any specific aspect of this global timekeeping system?

### **Nova Ai:**

The world clock goes from east to west due to the rotation of the Earth on its axis. This rotation causes the sun to appear to move across the sky from east to west, hence why time is traditionally referenced in that direction.

### **Deep Ai:**

The world clock, also known as the Earth's rotation, is a bit more complicated than just a simple clockwise or counterclockwise direction. The Earth's rotation is a complex process that involves the planet's angular momentum, its slightly ellipsoidal shape, and the gravitational interactions with other celestial bodies.



The Earth rotates on its axis from west to east, which means that the planet spins from left to right when viewed from above the North Pole. This is why the sun appears to rise in the east and set in the west. The rotation rate of the Earth is about 1,674 kilometres per hour (km/h) at the equator, which means it takes the planet about 24 hours to complete one full rotation.

Now, you might be wondering why the Earth rotates in this direction. There are a few reasons:

1. **Angular momentum:** When the solar system formed, the Earth and other planets started rotating due to the conservation of angular momentum. Angular momentum is a measure of an object's tendency to keep rotating or revolving around a central point. As the planets formed, they inherited some of this momentum from their early days as a swirling disk of gas and dust.
2. **Gravitational interactions:** The gravitational forces between the Earth and other celestial bodies, such as the Sun and other planets, play a significant role in shaping its rotation. The Sun's gravity causes the Earth to bulge out at the equator due to centrifugal force, which in turn affects its rotation.
3. **Core-mantle differentiation:** The Earth's internal structure is composed of a solid iron-rich core and a liquid iron-rich outer core. The movement of molten iron in the outer core generates electrical currents, which in turn produce magnetic fields. These magnetic fields interact with the Earth's rotation, influencing its angular momentum and direction.
4. **Planetary migration:** During the early days of the solar system, the planets migrated towards their current orbits due to gravitational interactions with each other and the Sun. This migration likely influenced the Earth's rotation rate and direction.

In summary, the world clock's direction is determined by a combination of factors, including angular momentum, gravitational interactions, core-mantle differentiation, and planetary migration.