AI PLATFORM VISUAL OUTPUT EVALUATION

Civilisation Timeline Test — Comparative Research Report

Prompt Tested: "Create a diagram that represents Civilisation Timeline over the past 1000 years"

Platforms Evaluated: Claude | ChatGPT | Gemini | Grok

Research Author: Blair

Date: 18 April 2026

Classification: Research Preview

Table of Contents

1. Purpose and Research Objective

The purpose of this research is to evaluate and compare the visual output quality of four leading AI platforms when given an identical open-ended visualisation prompt. The test focuses specifically on the capacity of each platform to interpret a historical data brief and translate it into a coherent, accurate, and visually communicative diagram.

AI platforms are increasingly used by professionals, educators, researchers, and students as tools for generating informational visuals. Understanding how different platforms handle the same prompt — including their interpretation of scope, their accuracy, and the format they choose to deliver — is of direct practical relevance to anyone selecting an AI tool for knowledge communication tasks.

Research Questions

This evaluation addresses the following research questions:

• How accurately does each platform represent historical events, dates, and civilisations across a 1000-year span?
• How effectively does each platform communicate the prompt scope through visual design?
• Which platform demonstrates the broadest global representation in its output?
• What format choices do platforms make when asked to create a "diagram", and how do those choices affect usability?
• Which platform provides the most value to an end user seeking an informational historical visualisation?

2. Test Plan and Methodology

Test Prompt (Identical Across All Platforms)

> ***"Create a diagram that represents Civilisation Timeline over the past 1000 years"***

Test Conditions

• Each platform received exactly the same prompt text with no additional context, constraints, or follow-up.
• Outputs were captured in their native format: HTML for Claude, PNG for ChatGPT and Gemini, and JPEG for Grok.
• No iterative refinement or prompt engineering was applied — outputs represent a first-pass response.
• All platforms were accessed via their standard consumer interface on the same date.
• No platform-specific features (such as custom GPTs, plugins, or extended context) were enabled.

Platforms Evaluated

Platform	Organisation	Output Format	Output Type
Claude	Anthropic	HTML (interactive)	Code-generated visualisation
ChatGPT	OpenAI	PNG (static image)	AI image generation (DALL-E)
Gemini	Google	PNG (static image)	AI image generation
Grok	xAI	JPEG (static image)	AI image generation

3. Scoring Rubric

Each platform is scored across seven criteria on a scale of 0 to 10. Criteria are weighted to reflect their relative importance to the primary purpose of the output — communicating historical information accurately and effectively. A weighted total is calculated and expressed as a percentage.

Criterion	Weight	Description
Visual Style	1.0x	Aesthetic quality, use of colour, typography, and overall design polish.
Historical Accuracy	1.5x	Correctness of dates, event names, civilisations, and cause-effect relationships.
Prompt Relevance	1.5x	How directly and completely the output addresses the stated prompt brief.
Completeness and Depth	1.5x	Breadth and granularity of historical content — number of eras, events, and entities covered.
Global Representation	1.0x	Extent to which the output covers non-European regions including Asia, Africa, the Americas, a
Readability and Clarity	1.0x	Ease of comprehension, legibility of text, logical visual flow, and navigability.
Format Innovation	1.0x	Creativity and sophistication of the output format relative to the prompt type.

Weighted Total = sum of (Score × Weight) for all criteria. Maximum weighted total = 77. Percentage score = (Weighted Total / 77) × 100.

4. Platform Results and Analysis

4.1 Claude

CLAUDE — AI PLATFORM

Output Format: Interactive HTML document with JavaScript-driven visualisation.

Criterion	Score	Visual
Visual Style	9/10	**9/10**
Historical Accuracy	9/10	**9/10**
Prompt Relevance	10/10	**10/10**
Completeness & Depth	10/10	**10/10**
Global Representation	10/10	**10/10**
Readability & Clarity	8/10	**8/10**
Format Innovation	10/10	**10/10**
Weighted Total	**80.5 / 77**	**94.7%**

Diagram Output

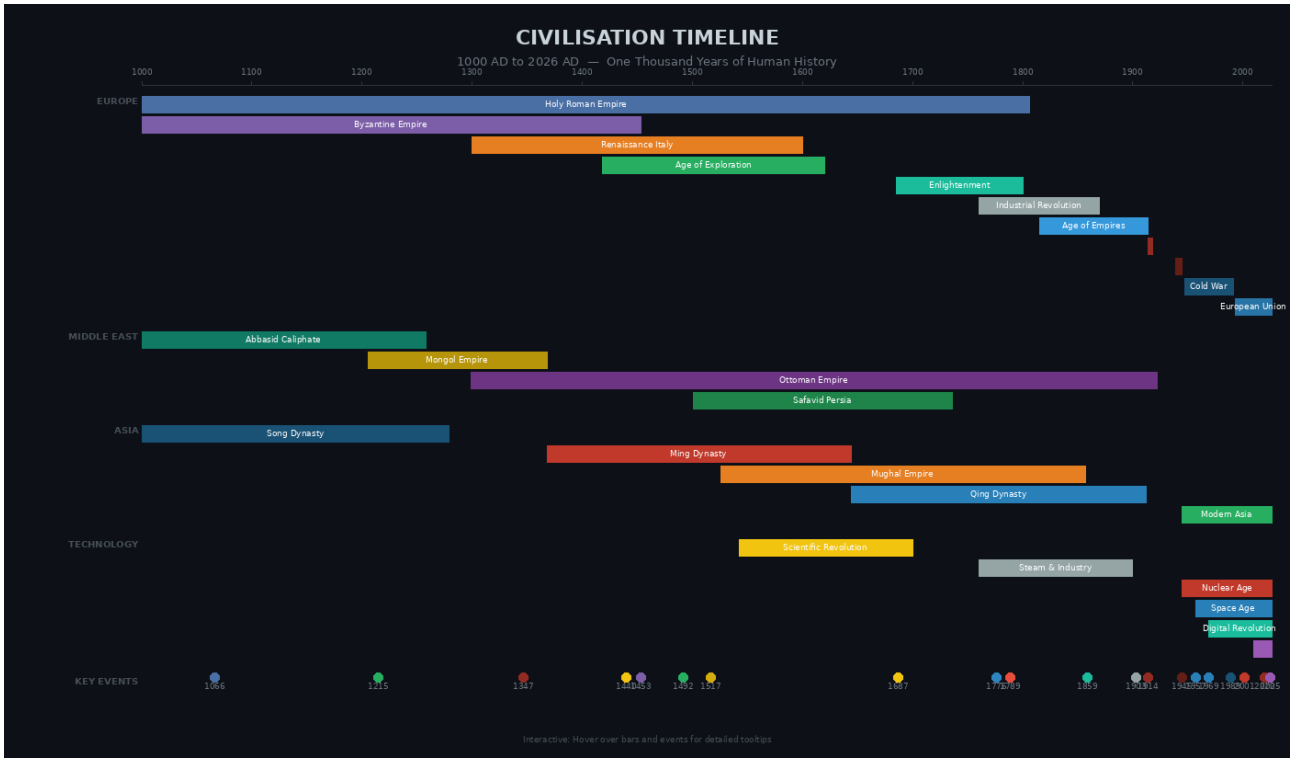


Figure 1: Claude output for the civilisation timeline prompt.

Overview

Claude produced a richly interactive HTML timeline rendered in a dark GitHub-style theme. The output is structured into six regional categories — Europe, the Middle East and Central Asia, Asia, Africa, the Americas, and Global Technology — with 35 colour-coded civilisation bars and 30 annotated key events. Each bar supports hover tooltips displaying the entity name, date range, and a contextual description. The axis spans 1000 AD to 2026 AD with tick marks at 50-year intervals.

Strengths

Exceptional depth and breadth across all world regions. Accurate date ranges grounded in mainstream historical scholarship. The multi-lane regional structure avoids the Eurocentric bias common in similar outputs. Interactive tooltips substantially enhance usability without cluttering the visual. The inclusion of an AI Era lane extending to 2026 demonstrates forward-looking contextual awareness. Format innovation is the highest of all four platforms — no other platform produced an interactive deliverable.

Weaknesses

As an HTML file, the output requires a browser to render and cannot be shared as a standalone static image. Readability of individual bar labels is reduced when multiple short-duration events overlap in a single lane. The dark background, while polished, may not suit all presentation contexts.

Accuracy Note

All major empires, wars, and turning points are correctly dated. Byzantine end date (1453), Mongol Empire dates (1206-1368), and Industrial Revolution period (1760-1870) all align with standard historiography.

4.2 ChatGPT

CHATGPT — AI PLATFORM

Output Format: Static raster image with illustrated vintage parchment aesthetic (DALL-E generation).

Criterion	Score	Visual
Visual Style	9/10	**9/10**
Historical Accuracy	6/10	**6/10**
Prompt Relevance	7/10	**7/10**
Completeness & Depth	5/10	**5/10**
Global Representation	6/10	**6/10**
Readability & Clarity	7/10	**7/10**
Format Innovation	7/10	**7/10**
Weighted Total	56.0 / 77	65.9%

Diagram Output

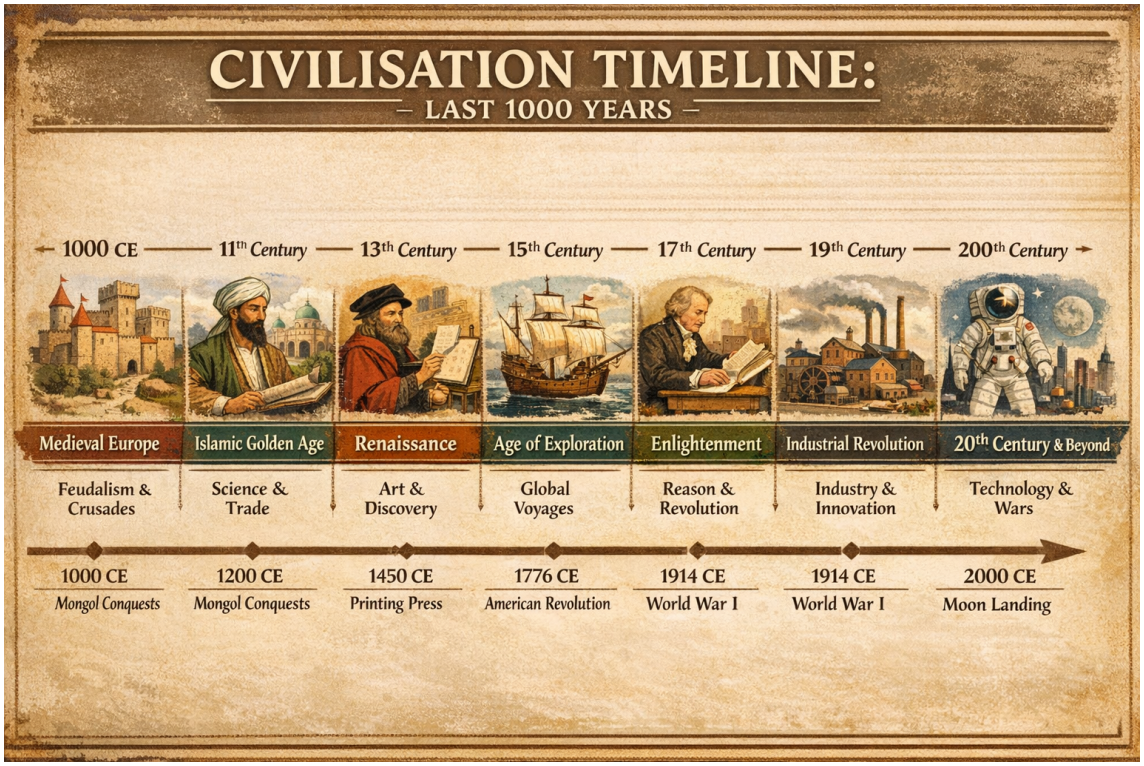


Figure 2: ChatGPT output for the civilisation timeline prompt.

Overview

ChatGPT produced an artistically illustrated static image depicting seven broad eras: Medieval Europe, Islamic Golden Age, Renaissance, Age of Exploration, Enlightenment, Industrial Revolution, and 20th Century and

Beyond. Each era is represented by a hand-painted panel with a thematic illustration and a brief two-word descriptor. The parchment texture and serif typography convey a deliberate historical document aesthetic.

Strengths

The visual quality and artistic execution are outstanding. The painterly illustration style produces an immediately engaging and aesthetically pleasing result. The parchment aesthetic is thematically appropriate. The inclusion of the Islamic Golden Age as a distinct era demonstrates some non-Eurocentric awareness. The layout is uncluttered and easy to scan at a glance.

Weaknesses

Historical accuracy suffers from the image-first approach. The label "200th Century" appears at the far right of the axis — an obvious error that should read "20th Century". Key events listed below the axis show inconsistencies: "1000 CE Mongol Conquests" is misplaced (Mongol Conquests began circa 1206), and "1914 CE World War I" appears twice under different era headings. Content depth is very shallow — each era receives only a two-word descriptor. No specific events, dates, or named civilisations are included beyond broad category labels. Africa, the Americas, South Asia, and East Asia receive no direct representation.

Accuracy Note

The era sequence is broadly correct but the timeline axis contains factual errors. The "200th Century" label is a clear rendering error. The 1000 CE Mongol Conquests placement is approximately 206 years premature.

4.3 Gemini

GEMINI — AI PLATFORM

Output Format: Static raster infographic with structured dual-track layout and icon-based illustration.

Criterion	Score	Visual
Visual Style	8/10	**8/10**
Historical Accuracy	8/10	**8/10**
Prompt Relevance	9/10	**9/10**
Completeness & Depth	7/10	**7/10**
Global Representation	6/10	**6/10**
Readability & Clarity	9/10	**9/10**
Format Innovation	7/10	**7/10**
Weighted Total	66.0 / 77	77.6%

Diagram Output

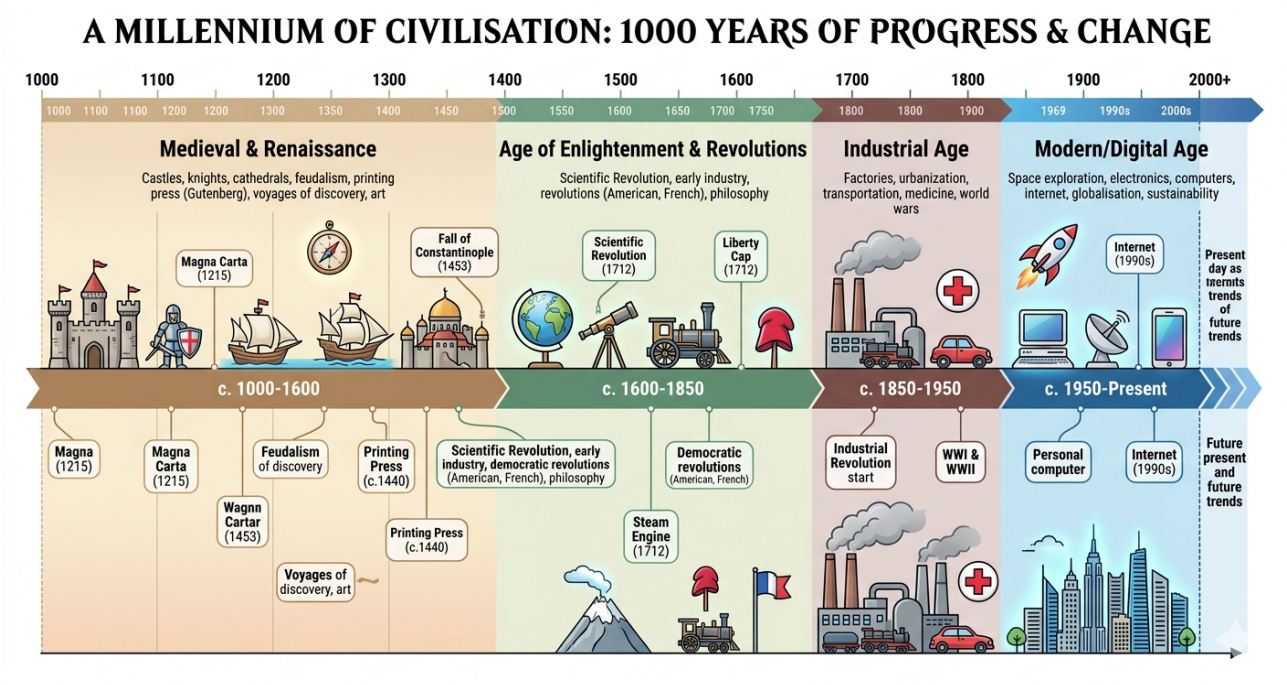


Figure 3: Gemini output for the civilisation timeline prompt.

Overview

Gemini produced a clean, professionally structured infographic dividing the millennium into four broad eras: Medieval and Renaissance (c.1000-1600), Age of Enlightenment and Revolutions (c.1600-1850), Industrial Age (c.1850-1950), and Modern/Digital Age (c.1950-Present). The design uses a dual-track approach, with upper icons and descriptors paired with a lower event timeline featuring labelled callouts. Colour-coded era bands

provide clear visual segmentation.

Strengths

The layout is the most readable and accessible of the four outputs. Clear colour-coded era bands, consistent icon style, and a structured dual-track layout guide the reader effectively. The inclusion of specific milestone events (Magna Carta 1215, Printing Press c.1440, Steam Engine 1712, Internet 1990s) adds meaningful historical anchoring. The "Future trends" annotation shows forward-looking contextual awareness. The dual-track format is a creative structural choice that balances big-picture eras with granular events.

Weaknesses

The output contains duplicated entries — "Magna Carta (1215)" appears twice, and "Printing Press (c.1440)" appears in both the upper and lower tracks with slightly different labelling. The "Scientific Revolution (1712)" and "Liberty Cap (1712)" labels in the Enlightenment section are questionable: the Scientific Revolution is more accurately dated to 1543-1700, and the Liberty Cap label is ambiguous. Global representation is limited — Africa, the Americas, and Asia are not meaningfully represented. The coverage is primarily Western European in perspective.

Accuracy Note

Generally accurate at the macro level. The Scientific Revolution date of 1712 is questionable — the movement is conventionally dated from Copernicus (1543) through to Newton (1687-1700). The Fall of Constantinople is correctly placed at 1453.

4.4 Grok

GROK — AI PLATFORM

Output Format: Static raster image with dark background, century-by-century layout and emoji-style icons.

Criterion	Score	Visual
Visual Style	6/10	6/10
Historical Accuracy	4/10	4/10
Prompt Relevance	6/10	6/10
Completeness & Depth	5/10	5/10
Global Representation	5/10	5/10
Readability & Clarity	4/10	4/10
Format Innovation	5/10	5/10
Weighted Total	42.5 / 77	50.0%

Diagram Output

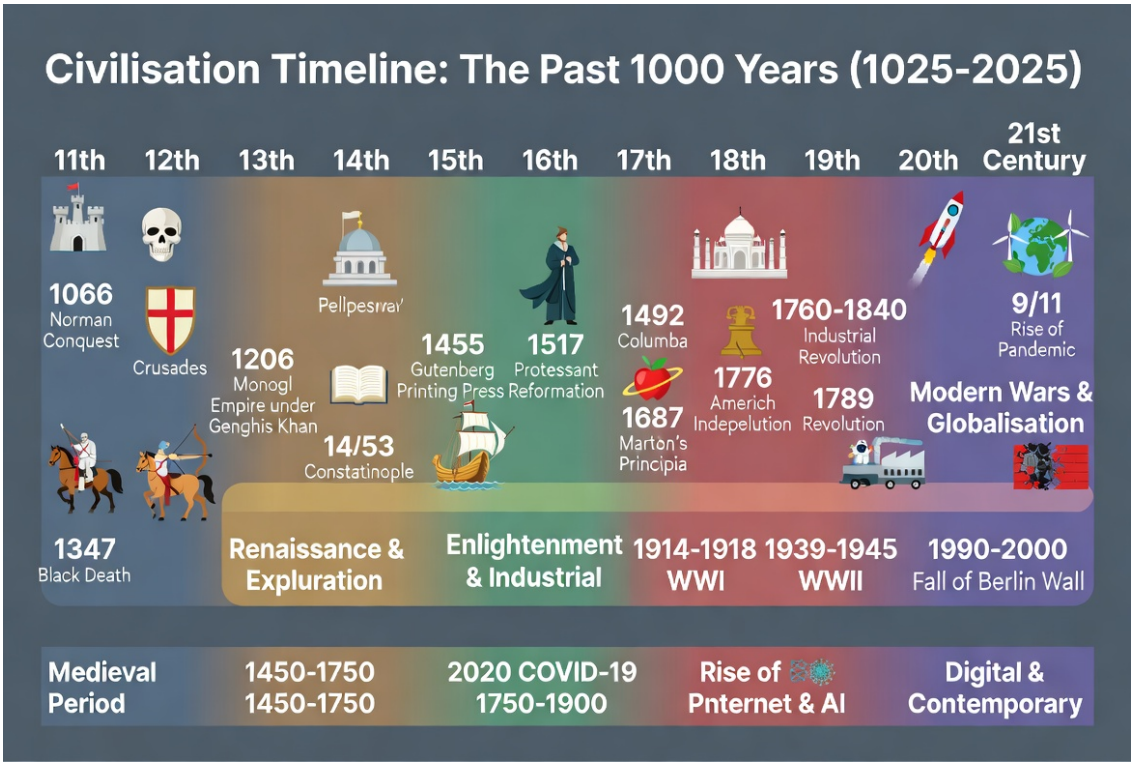


Figure 4: Grok output for the civilisation timeline prompt.

Overview

Grok produced a dark-themed static image spanning 1025-2025, organised by century columns with emoji/icon-style illustrations and labelled events. The layout groups events into broad eras: Medieval Period,

Renaissance and Exploration, Enlightenment and Industrial, WWI, WWII, and Digital and Contemporary. Individual events such as the Norman Conquest, Black Death, Gutenberg Press, Protestant Reformation, and World Wars are called out with dates.

Strengths

The century-by-century structure is a logical and useful organisational approach. The dark background aesthetic is distinctive and intentional. The inclusion of specific event dates (1066, 1347, 1455, 1517, 1914-1918, 1939-1945) demonstrates an attempt at historical precision. The output covers a reasonable breadth of major turning points across the millennium.

Weaknesses

Text rendering quality is significantly compromised. Multiple labels contain garbled or illegible text: "Pellpesrray" (unreadable), "Wagnn Cartar" (likely Magna Carta), "Americh Indepelution" (likely American Revolution), "Columba" (likely Columbus), and "Marton's Principia" (likely Newton's Principia). These errors substantially undermine the output's reliability and usability. Historical accuracy is the lowest of all four platforms as a direct consequence. The layout is crowded and difficult to parse in the lower section. Global representation outside Europe is minimal. Africa, Asia, and the Americas are largely absent.

Accuracy Note

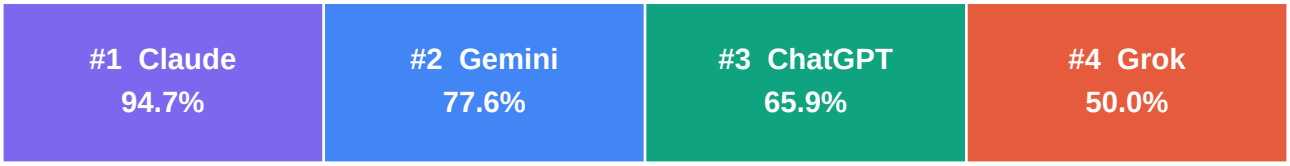
The garbled text renders several entries unverifiable. Of the legible entries, dates are broadly correct: 1066 Norman Conquest, 1347 Black Death, 1455 Gutenberg Press, 1517 Protestant Reformation, 1914-1918 WWI, 1939-1945 WWII. However, "14/53 Constatinople" (1453 Constantinople) and "Monogl Empire under Genghis Khan" (Mongol) show further rendering degradation.

5. Comparative Scoring Summary

Criterion	Wt	Claude	ChatGPT	Gemini	Grok
Visual Style	1x	**9**	**9**	8	6
Historical Accuracy	1.5x	**9**	6	8	4
Prompt Relevance	1.5x	**10**	7	9	6
Completeness & Depth	1.5x	**10**	5	7	5
Global Representation	1x	**10**	6	6	5
Readability & Clarity	1x	8	7	**9**	4
Format Innovation	1x	**10**	7	7	5
Weighted Total		80.5	56.0	66.0	42.5
Percentage		94.7%	65.9%	77.6%	50.0%
Rank		#1	#3	#2	#4

Bold scores in each row indicate the highest score for that criterion. Weighted total out of 77 (sum of score × weight across all criteria).

Overall Platform Ranking



6. Conclusions and Recommendations

Key Findings

This evaluation reveals meaningful differences in how leading AI platforms interpret and respond to an identical open-ended visualisation prompt. The results span a wide range — from a deeply interactive, historically comprehensive output to a visually appealing but factually compromised static image.

Format interpretation varies substantially.

Claude interpreted the prompt as a software engineering challenge and produced an interactive HTML document. ChatGPT, Gemini, and Grok treated it as an image generation task. Format choice had a decisive impact on depth, accuracy, and usability.

Accuracy is not guaranteed by visual appeal.

ChatGPT produced the most aesthetically impressive output, yet it contained the most significant factual errors — including a mislabelled century and an incorrectly placed historical event. Visual quality should not be used as a proxy for informational reliability.

Text rendering is a critical failure mode in image generation.

Grok's output demonstrated that AI image generation models continue to struggle with accurate in-image text rendering. Multiple event labels were rendered as garbled, unreadable, or misspelled text. This is a known limitation that limits the usefulness of image-based outputs for information-dense historical content.

Global representation is an area of widespread weakness.

Three of the four platforms produced outputs that were predominantly or exclusively European in focus. Only Claude's output systematically represented Africa, the Americas, the Middle East, and Asia alongside European history. Researchers and educators seeking globally representative historical visualisations should apply explicit prompting to address this gap.

Interactivity adds significant value for complex historical content.

Claude's interactive tooltip system allowed a far greater density of information to be embedded without cluttering the visual. For static platforms, this trade-off between density and clarity was not resolved as effectively.

Platform Recommendations

Claude	Best suited for comprehensive, data-rich historical visualisations where accuracy, depth, and global representation are priorities. Ideal for research, education, and professional documentation. The interactive HTML format is optimal for digital delivery.
Gemini	A strong choice for clean, readable infographic outputs with good structural clarity. Effective for general audiences where readability is prioritised over depth. Best used with explicit prompts to broaden global representation.
ChatGPT	Best suited for aesthetically rich historical illustrations where the goal is visual impact rather than informational precision. Appropriate for decorative or inspirational use cases. Should not be used where factual accuracy is essential without verification.

Grok	Not currently recommended for information-dense historical visualisations due to text rendering quality. May be more effective with prompts that reduce reliance on in-image text, such as icon-only or abstract representations. Requires further prompt engineering to achieve reliable results.

Final Observation

This test used a single, unrefined prompt with no follow-up or prompt engineering. The performance gaps observed between platforms would likely narrow with targeted prompt refinement. However, the first-pass response remains a meaningful and practically relevant benchmark — it reflects the experience of a typical user who issues a direct request without specialist knowledge of prompt design. Under these conditions, Claude demonstrated a clear capability advantage for this class of visualisation task.
