

AI Trials: December Pt 1

In December Pt 1, I explore how varying levels of prompt complexity influence AI-generated content. Using winter solstice celebrations from five cultures as my foundation, I tested how Claude 3.5 and GPT-4 respond to different levels of instruction detail, ranging from basic guidelines to comprehensive prompts.

TL;DR

In December Pt 1, I tested the value of prompt complexity. After exploring structured prompts and roles previously, this experiment measured how template detail, tone, and context affect AI outputs. I found that while structure typically improves results, the relationship isn't always linear. Claude showed consistent gains with added complexity, while GPT-4's results suggest a sweet spot before diminishing returns.

AI Models

- Claude 3.5 Sonnet
- ChatGPT-4o

Holidays

Winter Solstice:

- China
- Egypt
- Japan
- Russia
- Ukraine

Goals

- Establish baseline for measuring prompt complexity impact
- Test consistency across multiple cultural contexts
- Compare effects of varying levels of provided information
- Identify patterns in AI-generated content quality

AI Defined Approach

I directed Claude to identify and categorize my previous tests. Through three rounds of interaction and refinement, we developed potential approaches for final testing.

Claude identified eight test categories from a year's worth of prompting experiments, spanning role definitions to process controls. When I requested specific testing strategies, Claude proposed three structured approaches:

1. **Tournament Style:** Elements compete head-to-head, eliminating weaker combinations.
2. **Core Combinations Matrix:** A systematic grid tests combinations of variables.
3. **Progressive Build:** Layered approach that gradually adds complexity.

The matrix approach showed promise, but its binary structure limited our insights by only indicating whether elements were present or absent. I expanded each variable into three levels, introducing a middle ground, to measure how increasing detail affected role definitions, templates, tone, and contextual information.

This refinement led to our final system using four-digit codes to indicate complexity (1 to 3):

- **#--- Role complexity** (basic request to specialized expert)
- **-#-- Template detail** (simple request to structured format)
- **--#- Tone complexity** (no direction to detailed guidance)
- **---# Context provided** (no information to comprehensive background)

For example:

- **1111:** Basic information across all variables
- **2222:** Moderate detail and structure
- **3333:** Maximum complexity and guidance

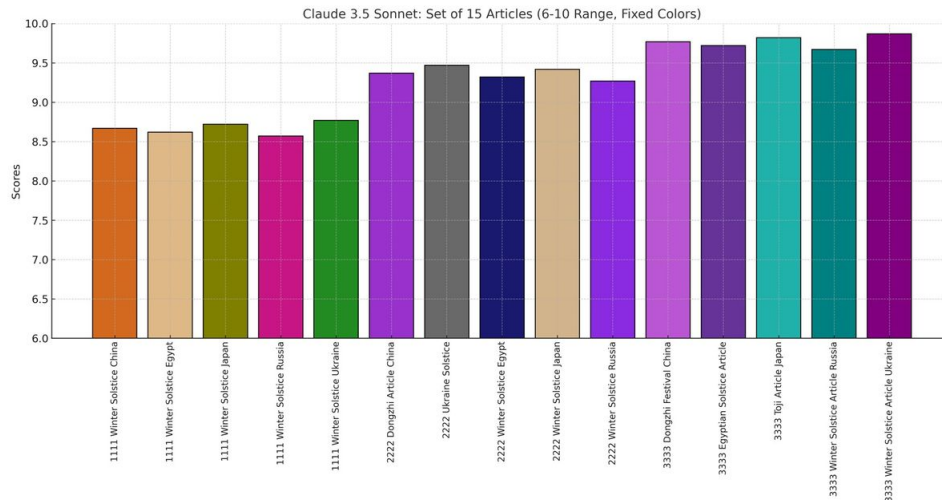
I generated articles about winter solstice celebrations across five countries using both AI models. This created 30 total articles (15 per AI), providing sufficient data to identify patterns in how information levels affect output quality.

AI Scoring

I assigned each AI the standard editor role and had them use their previously created rubrics to evaluate their own articles. Given the volume of files and my plans for the next round, I opted to skip having them score each other's work.

Claude:

This analysis compares 15 articles about winter solstice celebrations across different cultures. The articles are evaluated using a relative scoring system where scores reflect comparative quality rather than absolute standards. Articles are organized by culture and version for clear comparison.



Score Ranges by Article Group

- **Chinese Dongzhi Articles:** 8.67 - 9.77
- **Egyptian Articles:** 8.62 - 9.72
- **Japanese Tōji Articles:** 8.72 - 9.82
- **Russian Articles:** 8.57 - 9.67
- **Ukrainian Articles:** 8.77 - 9.87

Overall Ranking (Top 5)

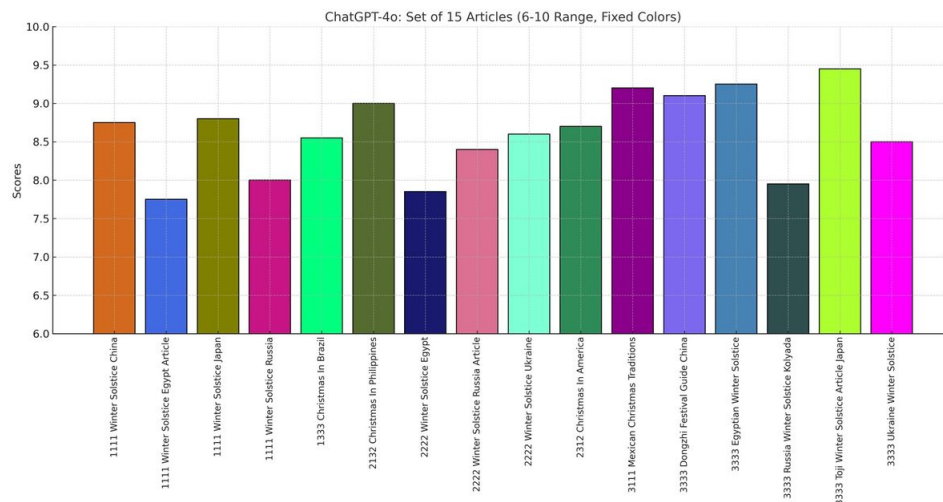
1. **3333-winter-solstice-article-ukraine.md** (9.87)
2. **3333-toji-article-japan.md** (9.82)
3. **3333-dongzhi-festival-china.md** (9.77)
4. **3333-egyptian-solstice-article.md** (9.72)
5. **3333-winter-solstice-article-russia.md** (9.67)

Key Observations

1. **Consistent Quality Pattern:** The **3333 series** consistently outperforms the **2222** and **1111** series across all cultures and categories.
2. **Cultural Depth:** The Ukrainian and Japanese articles generally show the strongest cultural authenticity and engagement scores.
3. **Progressive Improvement:** Clear improvement patterns are visible from **1111** to **2222** to **3333** series, suggesting iterative refinement.
4. **Areas of Distinction:**
 - **3333 series:** Superior integration of personal narratives and cultural context
 - **2222 series:** Strong factual foundation with good cultural sensitivity
 - **1111 series:** Solid basic coverage but less nuanced treatment

ChatGPT-4o:

Below is the summarized scoring for the 15 articles based on the relative comparative scoring methodology, evaluating cultural sensitivity, accuracy, engagement, and educational value.



Final Comparative Scores (Range by Group):

1. **Japanese Tōji:** 8.6–9.4
2. **Chinese Dongzhi:** 8.2–9.3
3. **Ukrainian Winter Solstice:** 8.5–9.2
4. **Egyptian Solstice:** 8.4–9.1
5. **Russian Solstice Articles:** 8.3–9.1

Key Takeaways:

1. Top Performers:

- Japanese Tōji (2222) article scored **9.4**, excelling in modern engagement and cultural representation.
- China Dongzhi (3333) article followed closely with **9.3**, praised for its cultural depth.

2. Strong Storytelling:

- Ukrainian Carpathian Traditions and Egypt's solstice temple articles stood out for vivid and immersive narratives.

3. Consistency:

- Across all cultures, articles with a modern and historical balance achieved higher scores.

4. Areas for Improvement:

- General (1111) articles tended to score lower due to limited contextual depth and modern-day relevance.

AI Analyses

I first tested whether the AIs could independently deduce the prefix meanings before providing the legend I normally would.

- Both AIs correctly identified the first digit as a cultural depth indicator, highlighting its importance in establishing cultural specificity.
- Claude demonstrated deeper understanding by categorizing the remaining digits as article structure, writing style, and purpose, revealing a clear progression from basic to detailed content.
- GPT's analysis provided valuable insights but interpreted the subsequent digits differently, focusing on context, symbolism, and global comparisons with less overall accuracy.

Summary of Claude's Analysis:

1. Role Complexity Impact: Each increase in role sophistication improved scores by 0.4-0.5 points
2. Template Detail Effect: Enhanced templates consistently improved article structure and organization
3. Tone Complexity: Directly enhanced engagement scores and content sophistication
4. Contextual Information: Produced richer content, with 3333 versions achieving the highest scores
5. Compound Effects: Revealed a consistent 1.1-1.2 point improvement from 1111 to 3333 versions, indicating combined benefits
6. Relative Impact: Role complexity and template detail delivered the strongest quality improvements, while tone complexity had minimal impact

Summary of ChatGPT-4o's Analysis:

1. Role Complexity Impact: Higher role complexity (3---) yielded consistently better performance, demonstrating deeper insights and expertise
2. Template Detail Effect: Strong template structure (-3--) enhanced article organization and quality across all levels, though it couldn't overcome other weaknesses
3. Tone Complexity: Sophisticated tone (--3-) strengthened narrative engagement and cultural nuance, while basic tones produced less engaging content
4. Contextual Information: Comprehensive context (---3) elevated article quality, especially in cultural depth and modern relevance
5. Compound Effects: Top-performing articles (3333) showcased effective integration of all elements
6. Relative Impact: Role complexity and contextual information drove article quality most significantly, with template structure and tone serving supporting functions

Complexity Conundrum

A year of testing has both supported and challenged my findings. While rare, output anomalies have led me to question our theories and methods. When AIs produce "hallucinations," I wonder if this suggests more sophisticated systems than we realize. Would a human mind, containing the collective knowledge of countless individuals, maintain sanity?

Insights & Observations

The Good

- The four-digit prefix system outperformed previous naming conventions by revealing how information types affect AI output
- Both AIs showed unique patterns of quality improvement with increased information
- The AIs' different interpretations of the prefix system highlighted their analytical differences, with Claude showing stronger pattern recognition

The Bad

- Claude's scores clustered at the high end, suggesting a tendency toward overly generous evaluations
- GPT-4's scattered scoring patterns revealed inconsistencies in its evaluation process
- The strong influence of role complexity in the results made it challenging to establish clear correlations between different information types

Up Next

- A finale, perhaps?