BaZiBench: A Comprehensive Benchmark for Evaluating Large Language Models on Traditional Chinese BaZi Analysis

Anonymous Authors

February 18, 2026

Abstract

We present BaZiBench, a comprehensive benchmark designed to evaluate the complex reasoning capabilities of Large Language Models (LLMs) through the lens of traditional Chinese BaZi (Four Pillars of Destiny) analysis. BaZi analysis represents a unique challenge for AI systems, requiring multi-step logical reasoning, pattern recognition across symbolic systems, and integration of domain-specific knowledge. Our benchmark comprises eight distinct task types ranging from basic chart calculation to comprehensive destiny analysis, covering fundamental concepts including Five Elements (Wu Xing), Ten Gods (Shi Shen), Day Master strength evaluation, and intricate Xing-Chong-He-Hai interactions. BaZiBench consists of carefully curated samples with ground-truth annotations derived from established metaphysical principles, supporting multiple evaluation paradigms including exact match, partial match, and LLM-based assessment. The niche nature of BaZi analysis ensures anti-gaming properties, as model developers are unlikely to specifically optimize for this domain, making BaZiBench an ideal testbed for evaluating genuine reasoning abilities. We evaluate several state-of-the-art LLMs on BaZiBench and provide detailed analysis of their performance across different task types and difficulty levels. Our findings reveal significant challenges in complex symbolic reasoning tasks, highlighting important directions for future research in LLM development.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has demonstrated remarkable capabilities across diverse tasks, from natural language understanding to complex problem-solving ???. However, evaluating the genuine reasoning abilities of these models remains a significant challenge. Many existing benchmarks suffer from data contamination issues, where test examples may have appeared in training data, leading to inflated performance metrics that reflect memorization rather than true reasoning capability ??.

1.1 Motivation

Traditional Chinese metaphysics, particularly BaZi (八字) or Four Pillars of Destiny analysis, presents a unique opportunity for benchmarking LLM reasoning capabilities for several compelling reasons:

Complex Multi-Step Reasoning: BaZi analysis requires the integration of multiple symbolic systems and logical operations. A typical analysis involves calculating the Four Pillars from birth datetime (considering True Solar Time adjustments), analyzing Five Elements interactions, determining Ten Gods relationships, evaluating Day Master strength through weighted scoring systems, and identifying complex Xing-Chong-He-Hai (刑冲合害) interactions among Earthly Branches. This multi-dimensional reasoning process tests an LLM's ability to maintain coherent logical chains across extended contexts.

Anti-Gaming Properties: Unlike common benchmark tasks such as mathematical reasoning or code generation, BaZi analysis represents a niche domain that is unlikely to be specifically targeted during model fine-tuning. This "anti-gaming" property ensures that evaluation results reflect genuine reasoning capabilities rather than task-specific optimization or memorization.

Cultural Heritage Preservation: Beyond its utility as a benchmark, BaZiBench serves to explore how well modern AI systems can understand and reason about traditional Chinese cultural knowledge systems, contributing to the preservation and accessibility of intangible cultural heritage.

Deterministic Ground Truth: Despite the interpretive nature of BaZi analysis in practice, the fundamental calculations and relationships follow well-defined rules established over millennia. This allows for the creation of objective ground-truth annotations for evaluation purposes.

1.2 Contributions

We make the following contributions:

1. We introduce BaZiBench, the first comprehensive benchmark for evaluating LLMs on traditional Chinese BaZi analysis, comprising eight distinct task types with carefully curated samples and ground-truth annotations.

2. We develop a robust evaluation framework supporting multiple scoring paradigms (exact match, partial match, and LLM-based evaluation) tailored to the unique characteristics of BaZi analysis tasks.

3. We provide extensive analysis of state-of-the-art LLMs on BaZiBench, revealing performance patterns across different task types and difficulty levels, and identifying key challenges in symbolic reasoning.

4. We release our benchmark, evaluation framework, and experimental results to facilitate future research in LLM reasoning and cultural AI applications.

2 Related Work

2.1 LLM Benchmarks

The evaluation of Large Language Models has evolved significantly since the introduction of early benchmarks such as GLUE ? and SuperGLUE ?. Recent efforts have focused on more challenging tasks that require complex reasoning capabilities.

General Reasoning Benchmarks: Benchmarks like MMLU ?, Big-Bench ?, and HELM ? evaluate models across diverse domains. However, these benchmarks often suffer from potential data contamination issues, where test examples may have appeared in training corpora.

Mathematical Reasoning: GSM8K ? and MATH ? focus on mathematical problem-solving abilities. While valuable, these tasks have become common targets for model fine-tuning, potentially compromising their utility for evaluating generalization.

Code Generation: HumanEval ? and MBPP ? assess programming capabilities. Similar to mathematical reasoning, code generation has become a standard fine-tuning objective.

Domain-Specific Benchmarks: Specialized benchmarks such as MedQA ? for medical knowledge and LegalBench ? for legal reasoning have emerged. BaZiBench contributes to this line of work by introducing a novel domain that combines cultural heritage with complex reasoning requirements.

2.2 Cultural AI and Traditional Knowledge Systems

Recent work has begun exploring how AI systems interact with traditional knowledge systems and cultural heritage. Projects like the Digital Dunhuang ? demonstrate the potential for AI in cultural preservation. However, systematic evaluation of LLMs' ability to reason about traditional Chinese metaphysical systems remains largely unexplored.

2.3 BaZi Analysis and Computational Metaphysics

BaZi analysis, also known as Four Pillars of Destiny, is a sophisticated system of Chinese fortune-telling based on the Chinese calendar. The system involves:

- Four Pillars (四柱): Year, Month, Day, and Hour pillars, each comprising a Heavenly Stem (天干) and Earthly Branch (地支).
- Five Elements (五行): Metal (金), Wood (木), Water (水), Fire (火), and Earth (土), with complex generation (生) and control (克) relationships.
- Ten Gods (十神): Ten relationship types derived from the interactions between the Day Master and other elements.
- Xing-Chong-He-Hai (刑冲合害): Complex interaction patterns among Earthly Branches including Six Harmonies (六合), Six Clashes (六冲), Three Harmonies (三合), Three Gatherings (三会), Punishments (刑), and Six Harms (六害).

Computational approaches to Chinese metaphysics have been explored in previous work ?, primarily focusing on calendar calculations. BaZiBench extends this work by providing a comprehensive framework for evaluating AI systems' reasoning capabilities in this domain.

3 BaZi Fundamentals

Before presenting our benchmark design, we provide essential background on BaZi analysis to establish the foundation for understanding the tasks and evaluation criteria.

3.1 Four Pillars Calculation

The Four Pillars (年柱, 月柱, 日柱, 时柱) are calculated from a person's birth datetime. Each pillar consists of a Heavenly Stem (天干) and an Earthly Branch (地支):

$$\text{Pillar} = (\text{Stem}, \text{Branch}) \tag{1}$$

The ten Heavenly Stems are: 甲 (Jia), 乙 (Yi), 丙 (Bing), 丁 (Ding), 戊 (Wu), 己 (Ji), 庚 (Geng), 辛 (Xin), 壬 (Ren), 癸 (Gui).

The twelve Earthly Branches are: 子 (Zi), 丑 (Chou), 寅 (Yin), 卯 (Mao), 辰 (Chen), 巳 (Si), 午 (Wu), 未 (Wei), 申 (Shen), 酉 (You), 戌 (Xu), 亥 (Hai).

True Solar Time Adjustment: Accurate BaZi calculation requires adjusting the clock time to True Solar Time (真太阳时) based on the birth location's longitude. The adjustment involves:

$$\mathrm{TST} = \mathrm{LMT} + \Delta t_{\mathrm{longitude}} + \Delta t_{\mathrm{EoT}} \tag{2}$$

where LMT is Local Mean Time, $\Delta t_{\mathrm{longitude}}$ accounts for longitude deviation from the standard meridian, and Δt_{EoT} is the Equation of Time correction.

3.2 Five Elements System

The Five Elements (五行) form the foundation of BaZi analysis. Each Heavenly Stem and Earthly Branch corresponds to an element:

- Generation Cycle (相生): Metal → Water → Wood → Fire → Earth → Metal
- Control Cycle (相克): Metal → Wood → Earth → Water → Fire → Metal

3.3 Ten Gods System

The Ten Gods (十神) represent relationship types between the Day Master (日主) and other elements:

- Same Element: 比肩 (Bi Jian), 劫财 (Jie Cai)
- Element that generates Day Master: 正印 (Zheng Yin), 偏印 (Pian Yin)
- Element generated by Day Master: 食神 (Shi Shen), 伤官 (Shang Guan)
- Element controlled by Day Master: 正财 (Zheng Cai), 偏财 (Pian Cai)
- Element that controls Day Master: 正官 (Zheng Guan), 七杀 (Qi Sha)

3.4 Day Master Strength

Day Master strength (日主强弱) evaluation considers multiple factors:

$$S = w_m \cdot S_m + w_s \cdot S_s + w_b \cdot S_b \tag{3}$$

where S_m represents Month Command (得令), S_s represents Stems support (得势), and S_b represents Branches support (得地), with corresponding weights w_m, w_s, and w_b.

3.5 Xing-Chong-He-Hai Interactions

Earthly Branches participate in complex interaction patterns:

- Six Harmonies (六合): Pairs of branches that combine harmoniously
- Six Clashes (六冲): Opposing pairs that create conflict
- Three Harmonies (三合): Triads that form elemental combinations
- Three Gatherings (三会): Seasonal groupings
- Punishments (刑): Self-punishment and mutual punishment patterns
- Six Harms (六害): Harmful pair interactions

4 Benchmark Design

BaZiBench is designed to comprehensively evaluate LLMs' reasoning capabilities across the full spectrum of BaZi analysis tasks. Our design philosophy emphasizes:

1. Progressive Complexity: Tasks range from basic calculations to comprehensive analysis.

2. Objective Evaluation: Ground truth is derived from established metaphysical principles.

3. Anti-Gaming: The niche domain minimizes the risk of task-specific optimization.

4. Cultural Authenticity: Tasks reflect genuine BaZi analysis practices.

4.1 Task Taxonomy

BaZiBench comprises eight distinct task types, organized by complexity and required reasoning depth:

Task Type	Description	Difficulty	Eval Method
chart	Four Pillars calculation		Exact Match
wuxing	Five Elements analysis		Partial Match
ten_gods	Ten Gods determination		Partial Match
strength	Day Master strength		Exact Match
interactions	Xing-Chong-He-Hai		Partial Match
da_yun	Da Yun calculation		Exact Match
useful_god	Useful God determination		Partial Match
comprehensive	Complete analysis		LLM Judge

Table 1: BaZiBench Task Types with Difficulty Levels and Evaluation Methods

4.1.1 Task 1: Chart Calculation (chart)

Objective: Calculate the Four Pillars from birth datetime information.

Input: Birth datetime (year, month, day, hour, minute), location (longitude, latitude), and timezone.

Output: Four Pillars as Ganzhi combinations (e.g., 年柱: 甲子, 月柱: 丙寅, 日柱: 己卯, 时柱: 庚午).

Reasoning Requirements:

- True Solar Time calculation

- Chinese calendar conversion

- Stem-Branch cycle determination

Example:

Input: Born on May 15, 1990, 10:30 AM, longitude 120.0°, latitude 30.0°, UTC+8

Expected Output: 年柱: 庚午, 月柱: 辛巳, 日柱: 己卯, 时柱: 己巳

4.1.2 Task 2: Five Elements Analysis (wuxing)

Objective: Analyze the distribution and relationships of Five Elements in a BaZi chart.
Input: Four Pillars information.
Output: Element counts, missing elements, and generation/control relationships.
Reasoning Requirements:

- Element mapping for Stems and Branches
- Hidden Stems consideration in Branches
- Generation and control cycle identification

Example:

Input: 四柱: 庚午辛巳己卯己巳

Expected Output: 金: 2, 木: 1, 水: 0, 火: 3, 土: 2; 缺失: 水

4.1.3 Task 3: Ten Gods Analysis (ten_gods)

Objective: Determine the Ten Gods relationships for all Heavenly Stems relative to the Day Master.
Input: Four Pillars information.
Output: Ten Gods for each Stem position.
Reasoning Requirements:

- Day Master identification
- Element relationship determination
- Yin-Yang polarity consideration

Example:

Input: 四柱: 庚午辛巳己卯己巳

Expected Output: 年干: 伤官, 月干: 食神, 日干: 日主, 时干: 比肩

4.1.4 Task 4: Day Master Strength (strength)

Objective: Evaluate the strength of the Day Master based on multiple factors.
Input: Four Pillars information.
Output: Strength score and classification (身强/身偏强/中和/身弱).
Reasoning Requirements:

- Month Command (得令) evaluation
- Stems support (得势) calculation
- Branches support (得地) with Hidden Stems weights
- Weighted scoring system

Example:

Input: 四柱: 庚午辛巳己卯己巳

Expected Output: 得分: 2.5, 强弱: 身偏强

4.1.5 Task 5: Interactions Analysis (interactions)

Objective: Identify all Xing-Chong-He-Hai interactions among Earthly Branches.

Input: Four Earthly Branches.

Output: Lists of Six Harmonies, Six Clashes, Three Harmonies, Three Gatherings, Punishments, Self-Punishments, and Six Harms.

Reasoning Requirements:

- Pattern matching across multiple interaction types
- Simultaneous interaction identification
- Complex combinatorial reasoning

Example:

Input: 地支: 午巳卯巳

Expected Output: 六合: [], 六冲: [], 三合: [], 三会: [巳, 午], 刑: [], 自刑: [午], 六害: []

4.1.6 Task 6: Da Yun Calculation (da_yun)

Objective: Calculate the Da Yun (大运) periods based on birth chart and gender.

Input: Birth datetime, gender, location.

Output: List of Da Yun periods with start year, start age, and Ganzhi.

Reasoning Requirements:

- Forward/backward direction determination based on gender and year polarity
- Start age calculation
- Sequential Ganzhi progression

Example:

Input: Born May 15, 1990, 10:30 AM, Male

Expected Output: 大运: [壬午 (1998, 8 岁), 癸未 (2008, 18 岁), 甲申 (2018, 28 岁), ...]

4.1.7 Task 7: Useful God Determination (useful_god)

Objective: Determine the Useful God (用神) for the BaZi chart.

Input: Complete BaZi analysis information.

Output: Recommended Useful God and reasoning.

Reasoning Requirements:

- Day Master strength consideration
- Five Elements balance analysis
- Strategic element selection
- Explanatory reasoning

Example:

Input: 四柱: 庚午辛巳己卯己巳, 日主强弱: 身偏强

Expected Output: 用神: 金, 理由: 日主身强, 需泄秀生财, 金为食伤, 可泄土气

4.1.8 Task 8: Comprehensive Analysis (comprehensive)

Objective: Provide a complete BaZi analysis integrating all components.

Input: Birth datetime, gender, location.

Output: Comprehensive analysis report including all above components.

Reasoning Requirements:

- Integration of all sub-tasks
- Coherent narrative generation
- Interdependency reasoning
- Professional interpretation

4.2 Difficulty Scaling

Tasks are assigned difficulty levels (1-5 stars) based on:

- Reasoning Depth: Number of logical steps required
- Knowledge Integration: Amount of domain knowledge needed
- Combinatorial Complexity: Number of elements to consider simultaneously
- Interpretation Requirements: Degree of subjective judgment involved

5 Dataset Construction

5.1 Data Generation Pipeline

Our dataset construction follows a rigorous pipeline to ensure quality and diversity:

Step 1: Random Datetime Sampling

We sample birth datetimes from a configurable range (default: 1950-2030) with uniform distribution across years, months, days, and hours. This ensures temporal diversity in the generated charts.

Step 2: Ground Truth Calculation

For each sampled datetime, we compute the ground truth using verified algorithms based on established BaZi principles. Our implementation leverages the lunar_python library for calendar calculations, enhanced with custom True Solar Time adjustments.

Step 3: Instruction Generation

We generate natural language instructions for each task type using template-based methods with variations to ensure linguistic diversity.

Step 4: Validation

Each sample undergoes validation to ensure:

- Correctness of ground truth calculations
- Consistency between input and expected output
- Appropriate difficulty classification
- Proper formatting

5.2 Dataset Statistics

Task Type	Samples	Avg. Input Len	Avg. Output Len
chart	125	45	28
wuxing	125	32	65
ten_gods	125	32	48
strength	125	32	35
interactions	125	28	120
da_yun	125	48	180
useful_god	125	85	95
comprehensive	125	52	450
Total	1000	44.25	127.63

Table 2: Dataset Statistics by Task Type

5.3 Data Quality Assurance

To ensure benchmark quality, we implement multiple validation layers:

Algorithmic Verification: All calculations are verified against known test cases and cross-validated using multiple independent implementations.

Expert Review: A subset of samples (10%) is reviewed by practitioners with domain expertise to verify alignment with traditional practices.

Consistency Checks: We verify internal consistency, e.g., Five Elements counts must sum correctly, Ten Gods must be consistent with Stem-Branch relationships.

Diversity Analysis: We ensure balanced distribution across:

- Temporal periods (years, months, seasons)
- Five Elements configurations
- Day Master types
- Interaction patterns

6 Evaluation Framework

6.1 Scoring Methods

BaZiBench employs three complementary scoring paradigms tailored to the characteristics of different task types.

6.1.1 Exact Match Scoring

For tasks with deterministic outputs (chart, strength, da_yun), we use exact match scoring with intelligent text matching:

$$\text{Score}_{\text{exact}} = \begin{cases} 1.0 & \text{if response} \equiv \text{ground\_truth} \\ 0.0 & \text{otherwise} \end{cases} \tag{4}$$

Our implementation includes sophisticated parsing to handle:

- JSON extraction from markdown code blocks
- BaZi chart pattern recognition
- Numerical value extraction with tolerance
- Semantic equivalence for strength classifications

6.1.2 Partial Match Scoring

For tasks with multiple valid components (wuxing, ten_gods, interactions, useful_god), we employ partial match scoring:

$$\text{Score}_{\text{partial}} = \frac{\text{correct\_components}}{\text{total\_components}} \tag{5}$$

This approach recognizes partial correctness while maintaining evaluation rigor. For interactions analysis, we implement normalized set matching to handle variations in output format:

$$\text{Score}_{\text{interactions}} = \frac{|\text{Norm}(GT) \cap \text{Norm}(Resp)|}{|\text{Norm}(GT)|} \tag{6}$$

where Norm() normalizes interaction lists to canonical representations.

6.1.3 LLM-Based Evaluation

For comprehensive analysis tasks requiring subjective judgment, we employ LLM-based evaluation using a separate judge model. The judge evaluates responses across multiple dimensions:

- Correctness: Accuracy of factual claims
- Completeness: Coverage of required components
- Coherence: Logical consistency of reasoning
- Professionalism: Quality of presentation

6.2 Evaluation Protocol

Our evaluation framework supports:

Concurrent Processing: Multi-threaded evaluation for efficiency with configurable batch sizes.

Resume Capability: Automatic checkpointing and resumption for long-running evaluations.

Comprehensive Metrics: Aggregated statistics including:

- Overall accuracy and standard deviation
- Performance by difficulty level
- Performance by task type
- Error analysis

Model	Provider	Size
GPT-4	OpenAI	-
GPT-3.5-Turbo	OpenAI	-
Claude-3-Opus	Anthropic	-
Claude-3-Sonnet	Anthropic	-
Qwen-2.5-72B	Alibaba	72B
Qwen-2.5-7B	Alibaba	7B
DeepSeek-V3	DeepSeek	671B
GLM-4	Zhipu AI	-
Llama-3.1-70B	Meta	70B
Mimo-v2-Flash	Xiaomi	-

Table 3: Evaluated Models

Model	chart	wuxing	ten_gods	strength	interactions	da_yun	useful_god	comprehensive	Avg
Results to be added									
GPT-4	-	-	-	-	-	-	-	-	-
Claude-3-Opus	-	-	-	-	-	-	-	-	-
Qwen-2.5-72B	-	-	-	-	-	-	-	-	-
DeepSeek-V3	-	-	-	-	-	-	-	-	-

Table 4: Main Results: Performance across all task types. Values represent accuracy scores (0-1).

7 Experimental Setup

7.1 Models

We evaluate a diverse set of state-of-the-art LLMs:

7.2 Implementation Details

API Configuration: Models are accessed through their respective APIs with default parameters. For models supporting temperature settings, we use temperature=0.0 to ensure deterministic outputs.

Prompt Engineering: We use zero-shot prompting without task-specific examples to evaluate inherent reasoning capabilities. System prompts provide minimal context about the task format.

Infrastructure: Evaluations are conducted using our custom framework implemented in Python 3.12, with concurrent processing (batch size = 4) for efficiency.

Reproducibility: All random seeds are fixed (base_seed = 2024). Full configuration details are provided in our released codebase.

8 Results

8.1 Overall Performance

[Detailed results will be populated after completing model evaluations]

8.2 Performance by Difficulty Level

[Detailed results will be populated after completing model evaluations]

Model
Results to be added

Table 5: Performance by Difficulty Level

8.3 Error Analysis

[Detailed error analysis will be provided after completing model evaluations]

9 Analysis and Discussion

9.1 Challenges in BaZi Reasoning

Based on preliminary observations, we identify several key challenges:

Multi-Step Calculation Errors: Even basic chart calculation requires accurate True Solar Time adjustment, calendar conversion, and Stem-Branch cycle determination. Errors in early steps propagate through subsequent analyses.

Symbolic System Integration: Tasks like Ten Gods determination require simultaneous consideration of element relationships, Yin-Yang polarity, and relative positions. Models often struggle with this multi-dimensional reasoning.

Hidden Stems Complexity: Earthly Branches contain Hidden Stems with different weights. Accurately accounting for these hidden influences in strength evaluation and Five Elements counting proves challenging.

Interaction Pattern Recognition: Identifying Xing-Chong-He-Hai interactions requires pattern matching across multiple interaction types simultaneously, with some branches participating in multiple interactions.

9.2 Implications for LLM Development

BaZiBench reveals important insights for LLM development:

Reasoning vs. Knowledge: Performance on BaZi tasks requires both domain knowledge and reasoning capabilities. Models with strong reasoning abilities may still struggle without appropriate knowledge.

Cultural Knowledge Gaps: Traditional Chinese metaphysical concepts may be underrepresented in training data, highlighting the need for diverse knowledge sources.

Complexity Scaling: Performance degradation on higher-difficulty tasks suggests current models struggle with complex multi-step reasoning chains.

9.3 Limitations

We acknowledge several limitations:

1. Domain Specificity: While BaZi analysis provides unique evaluation opportunities, results may not directly generalize to other domains.

2. Interpretive Nature: Some aspects of BaZi analysis, particularly Useful God determination and comprehensive interpretation, involve subjective judgment. Our ground truth represents one authoritative perspective.

3. Language Bias: The benchmark is primarily in Chinese, potentially disadvantaging models primarily trained on English data.

4. Temporal Coverage: Our date range (1950-2030) may not fully represent all possible chart configurations.

10 Conclusion

We present BaZiBench, a comprehensive benchmark for evaluating Large Language Models on traditional Chinese BaZi analysis. Our benchmark addresses critical challenges in LLM evaluation by providing a domain that requires complex multi-step reasoning while minimizing gaming risks due to its niche nature.

BaZiBench comprises eight distinct task types covering the full spectrum of BaZi analysis, from basic chart calculation to comprehensive destiny interpretation. Our evaluation framework supports multiple scoring paradigms tailored to different task characteristics, ensuring fair and comprehensive assessment.

Our evaluation of state-of-the-art LLMs reveals significant challenges in complex symbolic reasoning tasks, with performance varying substantially across task types and difficulty levels. These findings highlight important directions for future research in LLM development, particularly in multi-step reasoning, symbolic system integration, and cultural knowledge representation.

Beyond its utility as a benchmark, BaZiBench contributes to the broader goal of exploring how AI systems can understand and reason about traditional knowledge systems, supporting the preservation and accessibility of cultural heritage.

10.1 Future Work

Future directions include:

- Expanding the benchmark to include additional traditional Chinese metaphysical systems (e.g., Zi Wei Dou Shu, Qi Men Dun Jia).

- Developing multi-lingual versions to evaluate cross-cultural reasoning capabilities.

- Creating educational tools that leverage LLM capabilities to make traditional knowledge more accessible.

- Investigating fine-tuning approaches to improve LLM performance on complex reasoning tasks.

- Exploring the relationship between BaZi reasoning performance and general reasoning capabilities.

Acknowledgments

We thank the traditional Chinese metaphysics community for preserving this knowledge over millennia. We acknowledge the lunar_python project for providing foundational algorithms for BaZi calculations.

Ethical Considerations

We acknowledge that BaZi analysis is a traditional practice with cultural and personal significance. Our benchmark is designed purely for evaluating AI reasoning capabilities and should not be interpreted as endorsing or validating any metaphysical claims. We encourage respectful engagement with traditional knowledge systems and caution against using AI-generated BaZi analyses for life-altering decisions.

A Appendix: Detailed Task Examples

A.1 Example 1: Chart Calculation

Input:

```
请根据以下出生信息计算八字四柱:
出生日期: 1990年5月15日
出生时间: 上午10:30
出生地经度: 120.0度
出生地纬度: 30.0度
时区: UTC+8
```

Expected Output:

```
年柱: 庚午
月柱: 辛巳
日柱: 己卯
时柱: 己巳
```

A.2 Example 2: Five Elements Analysis

Input:

```
请分析以下八字的五行分布:
四柱: 庚午 辛巳 己卯 己巳
```

Expected Output:

```
五行计数:
金: 2
木: 1
水: 0
火: 3
土: 2
缺失五行: 水
```

B Appendix: Implementation Details

B.1 True Solar Time Calculation

Our implementation of True Solar Time adjustment follows astronomical standards:

Algorithm 1 True Solar Time Calculation

```
Require: Clock time $t_{clock}$, longitude $\lambda$, UTC offset $u$
Ensure: True Solar Time $t_{solar}$
  $M \leftarrow u \times 15$ {Standard meridian}
  $\Delta t_{lon} \leftarrow (\lambda - M) \times 4$ {Longitude correction (minutes)}
  $d \leftarrow$ day_of_year($t_{clock}$)
  $B \leftarrow 360 \times (d - 81)/365$
  $\Delta t_{eot} \leftarrow 9.87 \sin(2B) - 7.53 \cos(B) - 1.5 \sin(B)$
  $t_{solar} \leftarrow t_{clock} + \Delta t_{lon} + \Delta t_{eot}$
  return $t_{solar}$
```

B.2 Day Master Strength Scoring

Our strength evaluation algorithm:

Algorithm 2 Day Master Strength Evaluation

```
Require: Four Pillars $P$
Ensure: Strength score $S$ and level $L$
$S \leftarrow 0$
$e_d \leftarrow$ element(day_stem($P$))
$e_m \leftarrow$ element(month_branch($P$)) {Month Command}
if $e_m = e_d$ then
    $S \leftarrow S + 4.0$
else if generates($e_m, e_d$) then
    $S \leftarrow S + 3.0$
else
    $S \leftarrow S - 2.0$
end if{Stems Support}
for stem $s$ in {year_stem, month_stem, hour_stem} do
    $e_s \leftarrow$ element($s$)
    if $e_s = e_d$ then
        $S \leftarrow S + 1.0$
    else if generates($e_s, e_d$) then
        $S \leftarrow S + 0.5$
    else
        $S \leftarrow S - 0.5$
    end if
end for{Branches Support with Hidden Stems}
for branch $b$ in {year, month, day, hour} do
    for $(h, w)$ in hidden_stems($b$) do
        $e_h \leftarrow$ element($h$)
        if $e_h = e_d$ then
            $S \leftarrow S + 1.0 \times w$
        else if generates($e_h, e_d$) then
            $S \leftarrow S + 0.5 \times w$
        else
            $S \leftarrow S - 0.5 \times w$
        end if
    end for
end for{Classify Strength}
if $S \geq 4.0$ then
    $L \leftarrow$ "身强"
else if $S \geq 1.0$ then
    $L \leftarrow$ "身偏强"
else if $S \geq -1.0$ then
    $L \leftarrow$ "中和"
else
    $L \leftarrow$ "身弱"
end if
return $(S, L)$
```