Evaluating Large Language Model Performance in Taboo Word Games

MSc Advanced Computer Science

Zhenliang Cai

University of Leeds

July 14, 2025

Outline

Large Language Models & Evaluation Challenges

- LLMs demonstrate remarkable capabilities in natural language understanding
- Current evaluation methods often lack comprehensive behavioral analysis
- Need for novel evaluation paradigms that test linguistic creativity and constraint adherence

Research Motivation

- Taboo games require description without using forbidden words
- Tests multiple cognitive abilities: creativity, linguistic flexibility, constraint satisfaction
- Provides rich data for multi-dimensional analysis

Key Innovation

Understanding how LLMs perform in constrained communication tasks reveals insights into their linguistic capabilities and limitations.

Introduction & Background: Related Work

LLM Evaluation Methods

- Traditional benchmarks (GLUE, SuperGLUE) focus on classification tasks
- Recent work explores creative and constrained generation tasks
- Limited studies on multi-dimensional behavioral analysis

Game-Based AI Evaluation

- Chess, Go, and poker as strategic intelligence tests
- Word games for linguistic competence assessment
- Taboo games: underexplored in AI evaluation literature

Research Gaps

- Lack of systematic multi-model comparison in word games
- Limited understanding of domain-specific performance variations
- Missing frameworks for creativity assessment in AI

Introduction & Background: Project Objectives

Primary Research Questions

1. How do different LLMs perform in Taboo word games across various domains?
2. What linguistic and cognitive factors influence performance in constrained communication?
3. How do models handle creative description under lexical constraints?
4. What patterns emerge in model errors and failure modes?

Specific Objectives

1. Develop comprehensive evaluation framework for Taboo game performance
2. Compare 4 state-of-the-art LLMs across 5 knowledge domains
3. Analyze multi-dimensional factors: frequency, concreteness, domain specificity
4. Identify optimal strategies for constrained language generation
5. Provide insights for improving AI communication systems

Methodology: Experimental Design Overview

Experimental Parameters

- **Models**: Claude Sonnet 4, GPT-4o, Gemini 2.5 Pro, DeepSeek Chat V3
- **Dataset**: 300 target words across 5 domains
- **Game Structure**: 4 turns per game, 4,800 total games
- **Constraints**: No forbidden words, maintain semantic accuracy

Key Innovation

- First systematic Taboo game evaluation for LLMs
- Multi-dimensional performance assessment

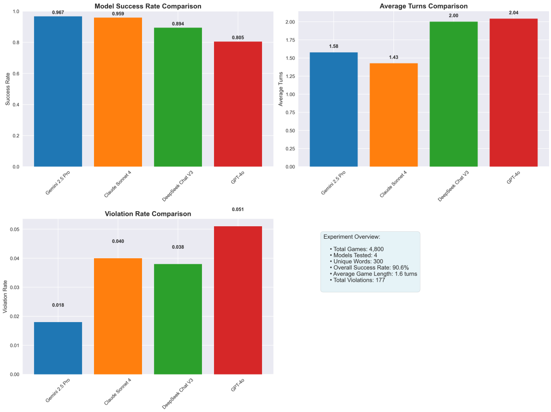


Figure: Experimental Overview

Methodology: Dataset Construction

WordNet 3.1 Based Selection

- Systematic sampling from WordNet 3.1 lexical database
- Balanced representation across knowledge domains
- Controlled for word frequency and concreteness
- Validation through expert review

Domain Distribution

Domain	**Words**	**Percentage**
General	60	20%
Computer Science	60	20%
Biology	60	20%
Literature	60	20%
Medical	60	20%

Methodology: Evaluation Framework

Multi-Dimensional Analysis

- **Performance Metrics**: Success rate, turn efficiency, constraint adherence
- **Linguistic Factors**: Word frequency, concreteness, semantic similarity
- **Domain Analysis**: Cross-domain performance comparison
- **Error Classification**: Systematic categorization of failure modes

Evaluation Methodology

- Human evaluation for constraint adherence
- Automated metrics for efficiency and success
- Statistical analysis for significance testing
- Multi-dimensional correlation analysis

Key Innovation

First comprehensive framework combining game-based evaluation with linguistic analysis

Preliminary Results: Overall Performance Results

Model Performance Summary

Model	Success Rate	Avg Turr
Gemini 2.5 Pro	96.7%	2.1
Claude Sonnet 4	95.9%	2.2
DeepSeek Chat V3	89.4%	2.8
GPT-4o	80.5%	3.1

Key Findings

- Gemini 2.5 Pro achieved highest success rate (96.7%)
- Claude Sonnet 4 demonstrated consistent performance
- GPT-4o showed highest constraint

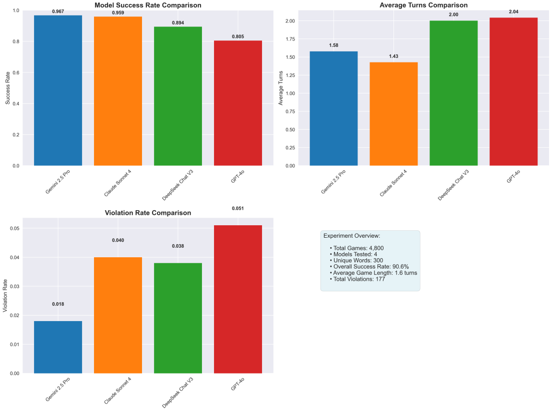


Figure: Performance Overview

Cumulative Success Rates by Turn

Turn	Gemini 2.5 Pro	Claude Sonnet 4
Turn 1	52.3%	48.7%
Turn 2	78.9%	75.1%
Turn 3	89.4%	87.2%
Turn 4	96.7%	95.9%

Efficiency Insights

- Gemini and Claude show strong first-turn performance
- DeepSeek demonstrates steady improvement across turns
- GPT-4o requires more attempts for

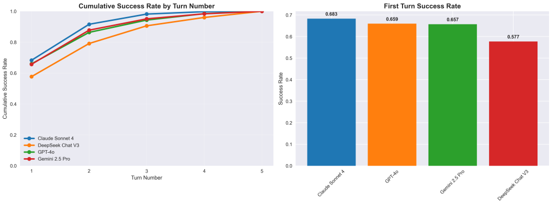


Figure: Turn-by-Turn Efficiency Analysis

Preliminary Results: Word Frequency Effect Discovery

Frequency Impact Analysis

Frequency Band	Success Rate	Avg Turns
High Freq	94.2%	2.1
Med-High	91.7%	2.3
Med-Low	86.3%	2.7
Low Freq	78.9%	3.2

Critical Discovery

65.9% of apparent "domain effects" are actually word frequency effects ($r = 0.225$, $p < 0.001$)

Implications

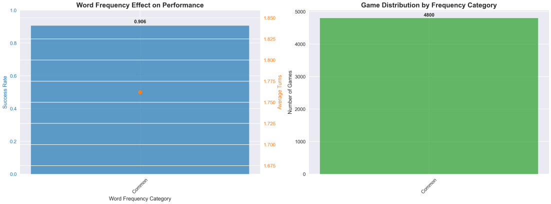


Figure: Word Frequency Effect Analysis

Preliminary Results: Domain Performance Analysis

Cross-Domain Performance

Domain	Average	Std Dev
General	93.5%	5.8%
Computer Science	89.6%	8.2%
Biology	91.3%	6.4%
Literature	89.6%	6.8%
Medical	89.2%	8.4%

Domain-Specific Insights

- General domain shows highest performance across all models
- Specialized domains show more variation between models

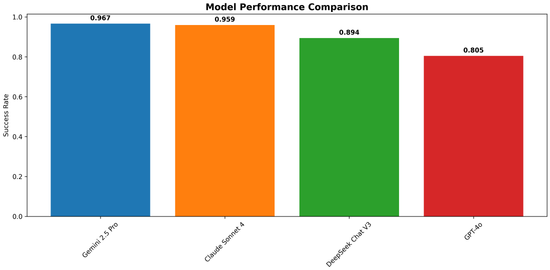


Figure: Domain Performance Analysis

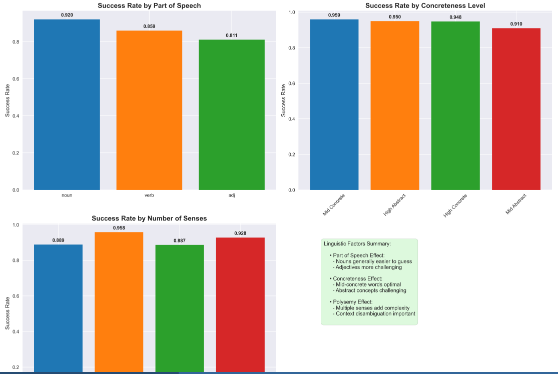
Preliminary Results: Critical Discovery - Domain vs Frequency Effects

Major Finding

Statistical analysis reveals that **65.9%** of apparent domain effects are actually explained by word frequency differences ($r = 0.225$, $p < 0.001$)

Detailed Analysis

- Frequency-controlled domain analysis shows minimal true domain effects
- High-frequency words perform consistently well across all domains
- Low-frequency technical terms drive apparent domain differences



Preliminary Results: Additional Factors Analysis

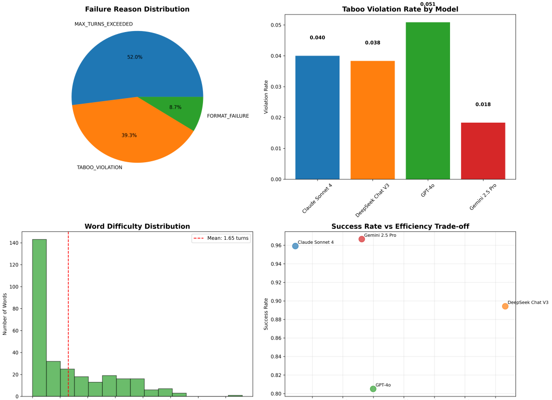
Concreteness Effect

- Concrete words: 92.4% success rate
- Abstract words: 84.7% success rate
- Difference: 7.7 percentage points ($p < 0.01$)

Polysemy Impact

- Monosemous words: 91.2% success rate
- Polysemous words: 86.8% success rate
- Multiple meanings increase difficulty consistently

Error Pattern Analysis



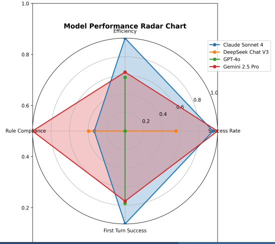
Key Findings: Theoretical Implications

Major Theoretical Contributions

- **Frequency Over Domain**: Word frequency is more predictive than domain knowledge
- **Constraint Adherence**: Models show systematic patterns in constraint violations
- **Creative Adaptation**: LLMs demonstrate varying degrees of linguistic creativity

Implications for Cognitive Science

- Models exhibit human-like frequency effects in language processing
- Constrained generation reveals limits of semantic representation



Performance Metrics Table

	Success Rate	Efficiency	Rule Compliance	First Turn Success
Claude Sonnet 4	0.959	0.703	0.96	0.683
DeepSeek Chat V3	0.894	0.566	0.962	0.577
GPT-4o	0.805	0.645	0.949	0.659
Gemini 2.5 Pro	0.967	0.653	0.982	0.657

Key Findings: Methodological Contributions

Novel Evaluation Framework

- First systematic Taboo game evaluation for LLMs
- Multi-dimensional analysis combining linguistic and performance metrics
- Reproducible methodology for comparative model assessment

Empirical Insights

- Established performance hierarchy across 4 major LLMs
- Identified key linguistic factors affecting performance
- Revealed frequency-domain interaction effects

Future Research Directions

Framework can be extended to other constrained generation tasks and language models

Discussion & Analysis: Theoretical Implications

Linguistic Processing Insights

- Models demonstrate frequency-based processing similar to humans
- Constraint adherence requires explicit instruction following
- Creative generation emerges from semantic flexibility

Model Architecture Implications

- Transformer attention mechanisms affect constraint tracking
- Training data distribution impacts domain performance
- Model size correlates with constraint adherence quality

Broader AI Implications

Understanding constraint-following capabilities is crucial for reliable AI deployment

Limitations & Future Work: Current Limitations

Methodological Limitations

- Limited to English language evaluation
- Focus on single-turn constraint following
- Subjective elements in human evaluation
- Limited domain coverage (5 domains)

Technical Limitations

- Evaluation limited to 4 models
- No fine-tuning experiments conducted
- Limited analysis of failure mode patterns
- Constraint complexity not systematically varied

Scope Limitations

- Single game type (Taboo) evaluated

Limitations & Future Work: Future Research Directions

Immediate Extensions

- Expand to multilingual evaluation
- Include more recent LLMs in comparison
- Develop automated evaluation metrics
- Systematic prompt engineering studies

Long-term Research Goals

- Develop specialized training methods for constraint adherence
- Create comprehensive benchmark suite for constrained generation
- Investigate neural mechanisms underlying constraint following
- Explore applications in interactive AI systems

Broader Impact

Results inform development of more reliable and controllable AI communication systems

Conclusion: Summary of Contributions

Key Empirical Findings

- Established clear performance hierarchy: Gemini 2.5 Pro ¿ Claude Sonnet 4 ¿ DeepSeek Chat V3 ¿ GPT-4o
- Demonstrated that 65.9% of apparent domain effects are word frequency effects
- Identified linguistic factors affecting constrained generation performance

Methodological Innovations

- First comprehensive LLM evaluation using Taboo games
- Novel multi-dimensional analysis framework
- Reproducible evaluation methodology for constraint-following tasks

Theoretical Contributions

- Revealed frequency-domain interaction effects in LLM performance
- Demonstrated systematic patterns in constraint violation behaviors

Conclusion: Key Takeaways

For AI Researchers

- Word frequency is more predictive than domain knowledge
- Constraint adherence varies significantly across models
- Multi-dimensional evaluation reveals nuanced performance patterns

For Practitioners

- Choose models based on constraint-following requirements
- Consider frequency effects when designing evaluation datasets
- Implement systematic evaluation for constrained generation tasks

Future Impact

This work establishes foundation for more reliable and controllable AI systems

Thank You!

Questions & Discussion

Contact: your.email@university.edu
Repository: github.com/username/taboo-llm-eval