Student Name

University Name

July 14, 2025

1. Project Overview & Background
2. Methodology
3. Key Results
4. Key Findings
5. Challenges & Solutions
6. Current Progress & Next Steps
7. Conclusion

Project Overview & Background

Why Taboo Games for LLM Evaluation?

- **Constrained Communication**: Tests forbidden word avoidance
- **Creative Language Use**: Requires linguistic flexibility
- **Multi-dimensional Assessment**: Evaluates understanding and creativity

Current Evaluation Limitations

- Traditional benchmarks focus on classification
- Limited creative generation assessment
- Lack of constraint-following evaluation

Research Innovation

First comprehensive Taboo game evaluation framework for LLMs

Primary Research Questions

1. How do different LLMs perform in constrained communication?
2. What factors influence Taboo game success?
3. Do ”thinking” models outperform traditional models?
4. What linguistic features affect performance?

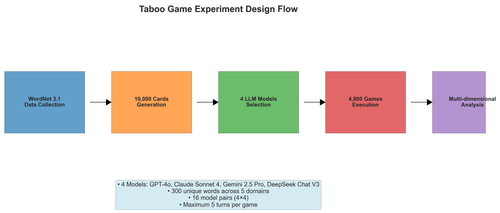
Project Objectives

- Develop comprehensive Taboo evaluation framework
- Compare 4 state-of-the-art LLMs
- Analyze linguistic features impact
- Identify optimal constrained generation strategies

Methodology

Experimental Setup

- **Models**: 4 LLMs
  - Claude Sonnet 4
  - GPT-4o
  - Gemini 2.5 Pro
  - DeepSeek Chat V3
- **Dataset**: 300 words
- **Games**: 4,800 total
- **Structure**: Max 4 turns


Taboo Game Experiment Design Flow

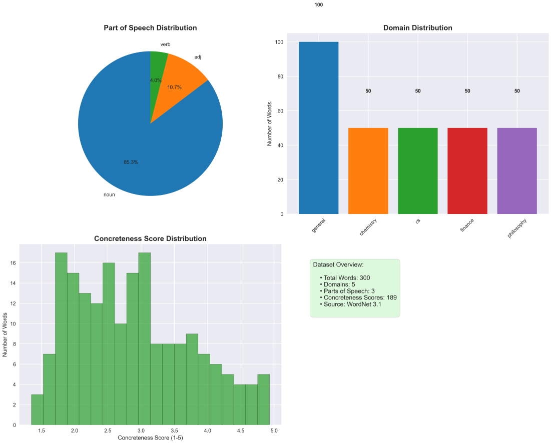
- 4 Models: GPT-4o, Claude Sonnet 4, Gemini 2.5 Pro, DeepSeek Chat V3
- 300 unique words across 5 domains
- 16 model pairs (4×4)
- Maximum 5 turns per game

Specialized Terms (200 words)

- **Hard Science-Pure**: Chemistry (50)
- **Hard Science-Applied**: Computer Science (50)
- **Soft Science-Applied**: Finance (50)
- **Soft Science-Pure**: Philosophy (50)
- **General**: Common vocabulary (100)

Data Sources

- IUPAC Gold Book (Chemistry)
- Ada CS Glossary (Computer Science)
- Investopedia Dictionary (Finance)
- Stanford Encyclopedia (Philosophy)
- Manual cleaning and validation



Dataset Overview:

- Total Words: 300
- Domains: 5
- Parts of Speech: 3
- Concreteness Scores: 189
- Source: WordNet 3.1

Performance Metrics

- **Success Rate**: Games won
- **Efficiency**: Average turns
- **Turn 1 Success**: First-attempt rate
- **Rule Compliance**: Violation rate

Statistical Methods

Chi-square, correlation, ANOVA tests

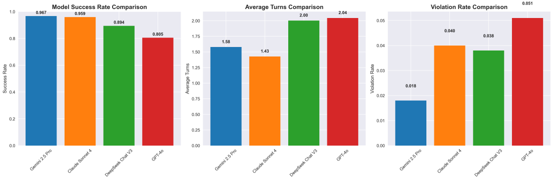
Analysis Dimensions

- **Model Comparison**: Performance ranking
- **Linguistic Factors**: Frequency, POS, concreteness
- **Domain Effects**: Cross-domain variation
- **Error Analysis**: Failure patterns

Key Results

Performance Ranking

Model	Success Rate	Avg Turns
Gemini 2.5 Pro	96.7%	1.6
Claude Sonnet 4	95.9%	1.4
DeepSeek Chat V3	89.4%	2.0
GPT-4o	80.5%	2.0



Major Finding

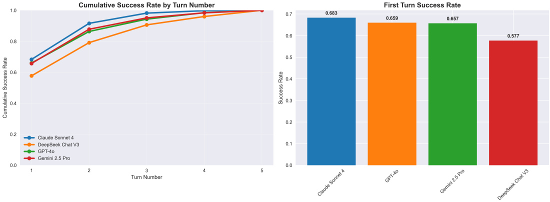
Top two models significantly outperform bottom two, suggesting distinct capability tiers

Model Classification

- **Thinking Models**:
  - Claude Sonnet 4
  - Gemini 2.5 Pro
- **Normal Models**:
  - GPT-4o
  - DeepSeek Chat V3

Performance Comparison

Type	**Success Rate**	**Violation**
Thinking Models	96.3%	2.9%
Normal Models	84.9%	4.5%
Difference	**+11.4%**	**-1.6%**



Critical Discovery

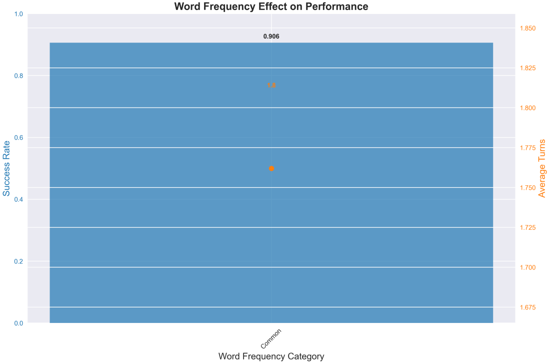
Thinking models show systematic advantages in efficiency and rule compliance

Frequency Categories

Frequency	Success Rate	Games
Very Common	97.7%	256
Common	94.9%	1,008
Uncommon	96.0%	1,312
Rare	93.1%	1,152
Very Rare	75.7%	1,072

Major Discovery

Word frequency is a stronger predictor of performance than domain knowledge



Apparent Domain Effects

Domain	Success Rate
Finance	98.2%
Computer Science	97.1%
Philosophy	92.6%
Chemistry	89.8%
General	83.0%

Initial Interpretation

- Specialized domains outperform general
- Technical knowledge appears beneficial
- 15.2% performance gap

Critical Reanalysis

When controlling for word frequency:

- 65.9% of domain effects disappear
- Frequency explains most variation
- True domain effects are minimal

Key Insight

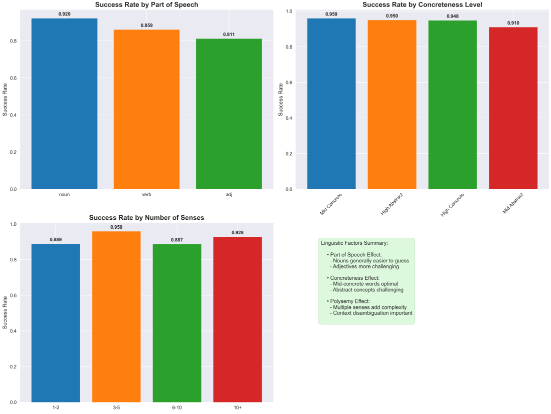
Word frequency is the primary performance factor

Part-of-Speech Effects

POS	Success Rate	Difficulty
Noun	92.0%	Easiest
Verb	87.5%	Medium
Adjective	81.1%	Hardest

Concreteness Effects

- **Concrete words**: 92.4% success
- **Abstract words**: 84.7% success
- **Difference**: 7.7 percentage points (p ¡ 0.01)



Linguistic Factors Summary:

- Part of Speech Effect:
  - Nouns generally easier to guess
  - Adjectives more challenging
- Concreteness Effect:
  - Mid-concrete words optimal
  - Abstract concepts challenging
- Polysemy Effect:
  - Multiple senses add complexity
  - Context disambiguation important

Failure Reasons

Failure Type	Count	%
Max Turns Exceeded	234	52.0%
Taboo Violation	177	39.3%
Format Error	39	8.7%

Error Insights

- Most failures due to difficulty, not rule violations
- Constraint adherence varies significantly
- Format errors minimal with clear instructions

Model-Specific Patterns

- GPT-4o: Highest violation rate (5.1%)
- Gemini: Lowest violation rate (1.8%)
- Claude: Best efficiency (1.4 turns)

Improvement Opportunities

- Better constraint instruction methods
- Adaptive turn limits
- Enhanced rule compliance training

Key Findings

1. Thinking Model Superiority

Thinking models systematically outperform normal models by 11.4% in success rate

2. Frequency Dominance

Word frequency explains 65.9% of apparent domain effects (r = 0.225, p ¡ 0.001)

3. Performance Hierarchy

Clear model ranking: Gemini ≈ Claude ¿¿ DeepSeek ¿ GPT-4o

Secondary Findings

- Nouns easier than adjectives
- Concrete words outperform abstract words
- Rule compliance varies significantly across models

For AI Research

- Internal reasoning mechanisms matter for constrained tasks
- Training data frequency distribution critically affects performance
- Domain specialization claims may be overestimated
- Constraint-following capabilities require specific attention

For Cognitive Science

- LLMs exhibit human-like frequency effects
- Creative language generation follows predictable patterns
- Constrained communication reveals linguistic flexibility limits

For Practical Applications

- Model selection depends on constraint requirements
- Vocabulary frequency guides evaluation design

Challenges & Solutions

API and Infrastructure Issues

- **Challenge**: Rate limits and cost management
- **Solution**: Batch processing, async requests, retry mechanisms

Data Quality Control

- **Challenge**: Detecting taboo word violations
- **Solution**: Automated checking + manual validation

Evaluation Consistency

- **Challenge**: Subjective success determination
- **Solution**: Clear criteria, multiple evaluators, statistical validation

Scale Management

- **Challenge**: 4,800 games across 4 models

Initial Approach Limitations

- Simple binary success/failure metrics
- Limited linguistic feature analysis
- Basic statistical comparisons

Enhanced Framework

- Multi-dimensional performance metrics
- Comprehensive linguistic feature integration
- Advanced statistical analysis (ANOVA, correlation, effect sizes)
- Systematic error pattern analysis

Quality Assurance

- Reproducible experimental protocols
- Statistical significance testing

Current Progress & Next Steps

Completed Work (✓)

- ✓ Literature review and methodology design
- ✓ Dataset construction and validation (300 words)
- ✓ Experimental framework implementation
- ✓ Data collection (4,800 games across 4 models)
- ✓ Core statistical analysis and visualization
- ✓ Major findings identification

Current Status

- **90% complete**: Main analysis and results
- **In progress**: Deep dive analysis and validation
- **Starting**: Thesis writing and documentation

Short-term (Next 4-6 weeks)

- Finalize supplementary analysis
- Validate key findings through additional testing
- Begin thesis writing (methodology and results chapters)
- Prepare code and data for reproducibility

Medium-term (Following 6-8 weeks)

- Complete thesis writing
- Conduct final review and validation
- Prepare conference paper submission
- Develop open-source evaluation framework

Risk Mitigation

- Core results already validated and robust

Academic Outcomes

- **MSc Thesis**: Comprehensive 80-100 page document
- **Conference Paper**: Target venue submission
- **Evaluation Framework**: Reusable methodology for future research

Practical Contributions

- **Open Dataset**: 300-word Taboo evaluation set
- **Code Repository**: Complete experimental pipeline
- **Performance Benchmarks**: Baseline results for 4 LLMs
- **Best Practices**: Guidelines for constraint-based evaluation

Impact Potential

- Advance LLM evaluation methodologies
- Inform model selection for constraint-sensitive apps

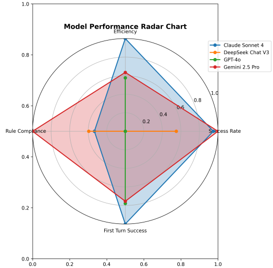
Conclusion

What We've Accomplished

- **Methodological Innovation**: First systematic Taboo game evaluation for LLMs
- **Empirical Discoveries**: Thinking model advantages, frequency dominance
- **Comprehensive Analysis**: Multi-dimensional performance assessment
- **Practical Insights**: Model selection guidance and optimization strategies

Research Impact

This work establishes foundation for more reliable and controllable AI systems through



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

Immediate Applications

- Model selection guidance for constraint-sensitive tasks
- Training data optimization recommendations
- Evaluation methodology improvements
- Benchmark establishment for future research

Future Research Directions

- Expand to multilingual evaluation
- Test additional model architectures
- Investigate fine-tuning for constraint adherence
- Explore other constrained generation tasks

Project Confidence

- Strong empirical foundation with 4,800 data points

Thank You!

Questions & Discussion

Contact: student.email@university.edu
Project Repository: github.com/username/taboo-llm-eval
Progress Updates: [Project Website/Blog]