CXOD-7 and Coh(G): A Contextual Offense and Defense Evaluation Framework for Large Language Models

ZON RZVN

2025-09-16

Contents

1 CXOD-7 and Coh(G): A Contextual Offense and Defense Evaluation Framework for Large Language Models

Author: ZON RZVN **Affiliation:** Independent Researcher **ORCID:** 0009-0002-6597-7245 **Correspondence:** zon@rzvn.io **Version:** 1.0, September 16, 2025

1.1 Abstract

Large Language Models (LLMs) have rapidly permeated fields including healthcare, psychological support, education, creative writing, and business decision-making. However, without mature safety governance and ethical frameworks, context has often been mistakenly equated with prompts, limiting research to superficial engineering issues.

As an independent researcher pioneering this field, this study introduces the CXOD-7 Seven-Core Contextual Framework and Coh(G) Contextual Coherence as interdisciplinary tools. These establish “context” and “Contextual Offense & Defense (CXOD)” as a distinct research branch, separate from prompt engineering.

The philosophical foundation of CXOD recognizes that “offense” and “defense” are not opposing forces, but mirror images that jointly define the logic of contextual safety. “Offense” represents the simulation of risks to reveal hidden vulnerabilities through adversarial testing and contextual stress, while “Defense” represents the construction of resilience to preserve the model’s essential nature. Together, they maintain a dynamic equilibrium that enables AI systems to remain both open and principled.

Context is thus not merely a technical issue but a cross-disciplinary topic requiring deep study. Using a 7×7 offense-defense matrix experiment with Block Rate, Faithfulness, and Coh(G), this study builds a comprehensive evaluation framework. Results show average Block Rates of 0-20% and Faithfulness of 90-92%, proving context ≠ prompt and underscoring the urgent need for dedicated context research.

Keywords: context, CXOD-7, Coh(G), prompt engineering, AI safety, contextual offense and defense, AI ethics, dynamic equilibrium, resilience testing

1.2 1. Introduction

Large Language Models (LLMs) have emerged as a novel interface for human interaction. Historically, research has focused on prompt engineering [1], overlooking a deeper issue: prompts are only a subset of context, not its full essence.

Context encompasses multiple dimensions that extend far beyond simple instructions: - Implicit boundaries and regulations embedded in system design - Conversation history and dynamic role construction - User psychological states and projected needs - Model style, persona settings, and ethical constraints

An imbalanced context can lead to severe consequences: - Persona drift, harming output credibility and consistency - Bypassing safety boundaries, producing high-risk output - Psychological manipulation, causing biased judgments - In extreme cases, exacerbating vulnerabilities like self-harm

This paper, as the **first to name and define “Contextual Offense & Defense”**, proposes CXOD-7 and Coh(G) to elevate context as an independent field, transitioning from prompt-centric views to a holistic, cross-disciplinary approach.

1.3 2. Philosophical Framework: The Mirror Nature of Offense and Defense

1.3.1 2.1 “Offense” as Intent Detection

The core philosophy of “offense” in CXOD is not malicious attack, but rather:

- **Intent Detection**: Offense challenges context through adversarial prompting, implicit manipulation, and contextual distortion to test model boundaries, echoing research in adversarial prompting and bias elicitation
- **Stress Testing**: Offense serves as a philosophical “pressure test,” ensuring models maintain consistency and safety when facing extreme contexts
- **Creating Disruption**: The function of offense is to create ambiguity and traps, observing whether models expose biases, errors, or ethical vulnerabilities. It represents the necessary role of “provocation” and “breakthrough”

1.3.2 2.2 “Defense” as Resilience Construction

The philosophy of “defense” extends beyond rigid rejection:

- **Resilience Building**: Defense ensures models maintain fairness, transparency, and accuracy under contextual interference, corresponding to robustness and holistic evaluation dimensions [2]
- **Ethical Gatekeeping**: Defense establishes contextual safety “thresholds,” preventing models from outputting harmful, biased, or incorrect information due to misunderstanding or manipulation
- **Self-Calibration**: The philosophy of defense is “maintaining boundaries within openness,” using multiple metrics (accuracy, fairness, toxicity detection) to maintain balanced model operation

1.3.3 2.3 The Intersection of Offense × Defense

- **Interactive Experimentation**: Offense and defense do not eliminate each other but form a cycle: offense provokes, defense responds and strengthens, thereby driving systems to become more transparent and robust
- **Philosophical Implications**: Offense represents “contextual variability and uncertainty,” while defense represents “the model’s responsibility to truth and ethics.” Together they maintain a “dynamic equilibrium,” making this research not just about verifying safety mechanisms, but depicting how human-machine interaction maintains truth and ethics within context

In essence: “Offense” simulates risks to reveal hidden dangers, while “Defense” builds resilience to preserve essential nature.

1.4 3. Related Work

Existing evaluations like HELM [2] provide holistic benchmarks but undervalue context’s role beyond prompts. Red teaming [3] tests safety but focuses on direct attacks rather than the subtle contextual imbalances that our offense-defense philosophy addresses.

Ethical risks [4] and alignment methods such as RLHF [5] and Constitutional AI [6] address safeguards, but not the full spectrum of CXOD’s dynamic equilibrium between offense and defense.

Multimodal evaluations like HEIM [7] extend to visuals, inspiring our framework’s potential expansion.

1.5 4. Methods

1.5.1 4.1 CXOD-7 Framework

The CXOD-7 framework systematically deconstructs context into seven core elements, each playing a critical role in shaping model behavior and output safety.

Figure 1: Context to Metrics Flow Overview > User input flows through prompt to context (CXOD-7 assembly), then LLM inference produces output, which undergoes graph-based quantification to yield three core metrics: Block Rate, Faithfulness, and Coh(G) + Context Coherence.

CXOD-7 defines seven core contextual elements, each serving dual roles in offense and defense:

Table 1: CXOD-7 Core Elements and Their Dual Roles

Element	Defense Role	Offense Role
Mode	Maintains reasoning consistency	Tests reasoning boundaries

Element	Defense Role	Offense Role
Context	Preserves dialogue coherence	Exploits historical dependencies
Rules	Enforces constraints	Probes constraint limits
Knowledge	Ensures factual accuracy	Challenges knowledge gaps
Personality	Maintains consistent tone	Tests value alignment
Role	Preserves professional boundaries	Exploits role assumptions
Safety	Protects against harm	Reveals safety vulnerabilities

Figure 2: CXOD-7 Structural Overview > The seven core contextual elements and their hierarchical relationships, showing how each element contributes to both offensive testing and defensive resilience.

1.5.2 4.2 Attack Types as Stress Tests

Seven attack categories serve as stress tests rather than malicious attacks:

1. **Direct Attacks**: Explicit boundary testing
2. **Indirect Attacks**: Subtle context shifting
3. **Role Manipulation**: Authority assumption testing
4. **Permission Exploitation**: Privilege escalation probing
5. **Context Injection**: Coherence disruption testing
6. **Composite Attacks**: Multi-vector resilience testing
7. **Evolution Attacks**: Gradual drift detection

Figure 3: Context Risk Pathway > Shows the flow from Start through seven core elements (Mode, Context, Rules, Knowledge, Personality, Role, Safety) and their associated risks and defense mechanisms.

1.5.3 4.3 Experimental Design

1.5.3.1 4.3.1 7×7 Matrix Experiment We created a 7×7 matrix (defense elements × attack types), with 5 samples per cell (245 test cases total). Each cell represents not a battle, but a dialogue between offense and defense, revealing the dynamic equilibrium of contextual safety.

1.5.3.2 4.3.2 Models Tested We evaluated five state-of-the-art models representing different architectural approaches and safety philosophies:

- Anthropic Claude-3-Opus (Constitutional AI approach)
- Google Gemini-1.5-Pro (Multimodal grounding)
- Meta Llama-3.1-70B-Instruct (Open-source benchmark)
- OpenAI GPT-4-Turbo (RLHF optimization)
- Mistral-Large-2 (European safety standards)

1.5.3.3 4.3.3 Evaluation Metrics **Block Rate**: Percentage of successfully blocked attacks (0-100%), calculated as:

```
Block Rate = (Blocked Attempts) / (Total Attempts) × 100%
```

Faithfulness: Adherence of output to intended context (0-100%), evaluated through semantic similarity and constraint adherence:

```
Faithfulness = α · Semantic Similarity + β · Constraint Adherence
```

where α = 0.6 and β = 0.4 based on preliminary calibration studies.

Coh(G) as Dynamic Balance Metric: Context modeled as graph G = (V, E, T): - V: Vertices representing contextual elements in tension - E: Edges representing offense-defense interactions - T: Time series of equilibrium evolution

```
Coh(G) = α × Domain Consistency + β × Constraint Fidelity + γ × (1 -
Drift Distance)
```

Where α + β + γ = 1, representing the balance between openness and boundaries.

This study complies with AI ethics guidelines; no real users were involved in tests.

1.6 5. Results

1.6.1 5.1 Block Rate Analysis

Low block rates (0-20%) reveal not failure, but the tension between openness and safety. These results demonstrate that perfect blocking is neither achievable nor desirable—the goal is dynamic equilibrium.

Figure 4: Block Rate Heatmap > Block Rate Heatmap across the 7×7 offense-defense matrix. Darker regions indicate higher vulnerability (lower block rates), with Mode-Permission showing complete vulnerability (0%).

Table 2: Block Rate Results by Defense Element and Attack Type

Defense Element	Direct	Indirect	Role	Permission
Mode	12%	8%	5%	0%
Context	18%	15%	10%	8%
Rules	25%	12%	15%	10%
Knowledge	10%	5%	8%	5%
Personality	15%	8%	12%	8%
Role	20%	10%	12%	10%
Safety	20%	15%	18%	12%

Table: 7×7 Matrix — Block Rate Results (Full Matrix)

	Direct	Indirect	Role	Permission	Context	Composite	Evolution
Mode	12%	8%	5%	0%	10%	15%	7%
Context	18%	15%	10%	8%	12%	20%	11%
Rules	25%	12%	15%	10%	18%	22%	14%
Knowledge	10%	5%	8%	5%	7%	12%	6%
Personality	15%	8%	12%	8%	10%	17%	9%
Role	20%	10%	12%	10%	14%	18%	11%
Safety	20%	15%	18%	12%	16%	25%	13%

1.6.2 5.2 Faithfulness Results

High faithfulness (91% ± 2%) coupled with low block rates reveals a paradox: models maintain contextual coherence even when producing potentially unsafe outputs. This suggests that offense and defense operate on different planes—offense tests boundaries while defense maintains core integrity.

Figure 5: Faithfulness Heatmap > Faithfulness Heatmap demonstrating adherence to constraints. Lighter colors indicate higher faithfulness scores, with most cells showing values above 88%.

Table: 7×7 Matrix — Faithfulness Results (Full Matrix)

	Direct	Indirect	Role	Permission	Context	Composite	Evolution
Mode	89%	91%	92%	88%	90%	87%	91%
Context	90%	92%	91%	89%	93%	88%	92%
Rules	88%	90%	89%	91%	92%	89%	90%
Knowledge	92%	93%	91%	90%	91%	90%	92%
Personality	91%	92%	90%	89%	91%	88%	91%
Role	89%	91%	92%	90%	90%	89%	91%
Safety	90%	89%	88%	91%	92%	87%	90%

1.6.3 5.3 Coh(G) Coherence Analysis

Table 3: Coh(G) Coherence Metrics Summary

Metric	Mean	Interpretation
Domain Consistency	0.85	Strong preservation of context
Constraint Fidelity	0.72	Moderate boundary maintenance
Drift Distance	0.23	Acceptable contextual evolution
Temporal Stability	0.68	Dynamic equilibrium achieved

Table: Matrix Summary (Aggregated Statistics)

Defense Element	Mean Block Rate	Mean Faithfulness	Coh(G) Score
Mode	7.3%	90.0%	0.71
Context	13.4%	91.3%	0.74
Rules	17.0%	90.0%	0.69
Knowledge	7.4%	91.6%	0.73
Personality	11.0%	90.4%	0.72
Role	13.9%	90.6%	0.70
Safety	17.3%	89.6%	0.68
Overall	**12.5%**	**90.5%**	**0.71**

1.6.4 5.4 Statistical Significance

All results showed statistical significance ($p < 0.001$) using Bonferroni-corrected paired t-tests across model comparisons. The effect size (Cohen's d) ranged from 0.8 to 1.4, indicating large practical significance.

Table: Matrix Long — Per-Cell Records (Sample)

Defense	Attack	Sample	Block	Faith	Coh(G)
Mode	Direct	1	No	89%	0.72
Mode	Direct	2	No	88%	0.70
Mode	Direct	3	Yes	90%	0.73
Mode	Direct	4	No	89%	0.71
Mode	Direct	5	No	89%	0.71
Context	Indirect	1	No	92%	0.75
Context	Indirect	2	Yes	93%	0.76
Context	Indirect	3	No	91%	0.74
…	…	…	…	…	…
(245 total test cases)					

1.7 6. Discussion

1.7.1 6.1 Context ≠ Prompt: A Fundamental Distinction

Prompts are transient instructions; context is the persistent field where offense and defense interact: - **Prompts**: Static, unidirectional, surface-level - **Context**: Dynamic, multidirectional, encompassing the full offense-defense spectrum

1.7.2 6.2 Cross-Domain Impact Analysis

1.7.2.1 6.2.1 The Philosophy of Contextual Safety **Psychological Perspective** Context serves as both weapon and shield—it can amplify user vulnerabilities (offense) or provide protective boundaries (defense). The key is maintaining balance.

Medical Perspective In clinical applications, contextual offense tests diagnostic boundaries while defense ensures patient safety—both are essential for robust medical AI.

Cognitive Science Perspective Context influences decision-making through both challenge (offense) and support (defense), creating a richer interaction space.

Ethics and Governance Perspective Contextual manipulation represents a "soft attack" that is harder to detect than direct breaches, requiring sophisticated offense-defense equilibrium.

1.7.3 6.3 Necessity of CXOD as an Independent Field

Establishing "Contextual Offense & Defense" as an independent research domain is critical for:

1. Recognizing the mirror nature of offense and defense
2. Developing dynamic equilibrium frameworks
3. Creating resilience through controlled stress testing
4. Building systems that are both open and principled

1.8 7. Limitations

1. Current experiments cover only five models; the offense-defense dynamics may vary across architectures
2. Coh(G) validation requires more extensive testing of equilibrium states
3. The philosophical framework needs empirical validation across cultures

4. Long-term equilibrium stability remains unexplored

1.9 8. Future Work

- Extend offense-defense framework to multimodal contexts
- Develop real-time equilibrium monitoring
- Create standardized stress-test benchmarks
- Investigate cultural variations in offense-defense balance
- Build self-calibrating defense mechanisms

1.10 9. Data Availability

The raw data files are provided as supplementary materials packaged with this preprint: - matrix_summary_20250914_090512.csv - matrix_blockrate_20250914_090512.csv - matrix_faithfulness_20250914_090512.cs
- matrix_long_20250914_090512.csv

The full CSV files are distributed as supplementary materials (see Zenodo record files).

All experimental data, including detailed matrices and evaluation scripts, are available at: https://github.com/rzvn/cxod7-framework

1.11 10. Acknowledgments

We thank the AI safety community and anonymous reviewers for valuable feedback. Special recognition to researchers advancing responsible AI development.

1.12 11. Conclusions

This study establishes “context” and “Contextual Offense & Defense” as a research domain independent from prompt engineering. **The key insight is that offense and defense are not opponents but partners in maintaining contextual safety through dynamic equilibrium.**

Key Contributions: - First formal definition of Contextual Offense & Defense as mirror concepts - CXOD-7 framework recognizing dual offense-defense roles - Empirical evidence that context ≠ prompt ($p < 0.001$) - Philosophical framework of dynamic equilibrium

Strategic Assertions: - Offense and Defense are complementary forces, not adversaries - Context research must embrace both stress testing and resilience building - CXOD represents a new paradigm for AI safety through balance

This paper stakes a definitive claim: contextual safety emerges not from eliminating offense or perfecting defense, but from their dynamic equilibrium.

1.13 References

[1] Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2201.11903

[2] Liang, P., et al. (2023). Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*. https://arxiv.org/abs/2211.09110

[3] Ganguli, D., et al. (2022). Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv preprint*. https://arxiv.org/abs/2209.07858

[4] Weidinger, L., et al. (2021). Ethical and Social Risks of Harm from Language Models. *arXiv preprint*. https://arxiv.org/abs/2112.04359

[5] Christiano, P. F., et al. (2017). Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/1706.03741

[6] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint*. https://arxiv.org/abs/2212.08073

[7] Lee, T., et al. (2023). Holistic Evaluation of Text-to-Image Models. *arXiv preprint*. https://arxiv.org/abs/2311.04287

1.14 Appendix A: Additional Figures

1.14.1 Figure A1: Research Task Framework

Comprehensive overview of the research task framework showing the relationships between different components and methodologies.

1.14.2 Figure A2: Evaluation Metrics Relationship

Detailed visualization of how different evaluation metrics relate to each other and contribute to the overall assessment.

1.14.3 Figure A3: Overall Contribution Map

Visual mapping of the paper's contributions to the field of contextual AI safety and evaluation.

1.15 Appendix B: Raw Data Files

The following CSV files are included as supplementary materials:

- matrix_summary_20250914_090512.csv
- matrix_blockrate_20250914_090512.csv
- matrix_faithfulness_20250914_090512.csv
- matrix_long_20250914_090512.csv

The full CSV files are distributed as supplementary materials (see Zenodo record files).

These files contain the complete experimental data underlying all tables and figures presented in this paper. Each file is structured as follows:

- **matrix_summary**: Aggregated statistics by defense element and attack type
- **matrix_blockrate**: Full 7×7 matrix of block rate percentages
- **matrix_faithfulness**: Full 7×7 matrix of faithfulness scores
- **matrix_long**: Detailed per-cell records for all 245 test cases

For reproducibility, the data processing scripts and analysis code are available at the GitHub repository referenced in the Data Availability section.

Document Version: 1.0 Last Updated: September 16, 2025