The 70% Rule: When Axiomatic Prompting Helps, and When It Hurts

Adam Zachary Wasserman
Independent Researcher
wassermana@gmail.com

January 2025

Abstract

We introduce *axiomatic prompting*, a technique that augments LLM classification prompts with explicit decision rules of the form *pattern* → *category*. Across six classification tasks and eight models spanning five providers (OpenAI, Anthropic, Google, Meta, Alibaba), we find a consistent threshold effect: axiomatic prompting significantly improves accuracy when zero-shot performance falls below 70% (gains of +15–24%, $p < 0.05$), but significantly *hurts* performance above this threshold (losses of -5 to -20%, $p < 0.05$). We term this the **70% Rule**, a simple heuristic that predicted the direction of axiomatic prompting's effect with 100% accuracy in our experiments. Ablation studies reveal that axiom style matters: pattern-based rules derived from data outperform conceptual definitions. Our findings suggest practitioners should evaluate zero-shot baselines before investing in rule-based prompt engineering. Code and data available at https://doi.org/10.17605/OSF.IO/PCX2D.

1 Introduction

When deploying large language models for classification, practitioners face a fundamental choice: trust the model's learned representations, or provide explicit rules to guide its decisions. Intuition suggests that domain-specific rules should help; after all, experts use rules, so why not teach them to the model?

We tested this intuition systematically and found it wrong—or rather, conditionally wrong.

The 70% Rule

Before adding classification rules to your prompt, run zero-shot first.

- **Below 70%**: Rules help. Expect +15–24% improvement.
- **Above 70%**: Rules hurt. Expect -5 to -20% degradation.

This rule predicted the correct direction in 100% of our experiments (6/6 tasks).

We introduce *axiomatic prompting*: providing LLMs with explicit classification axioms that map observable patterns to categories. An axiom takes the form:

```
IF text contains "governed by" OR "construed in accordance"
THEN classify as "Governing Laws"
```

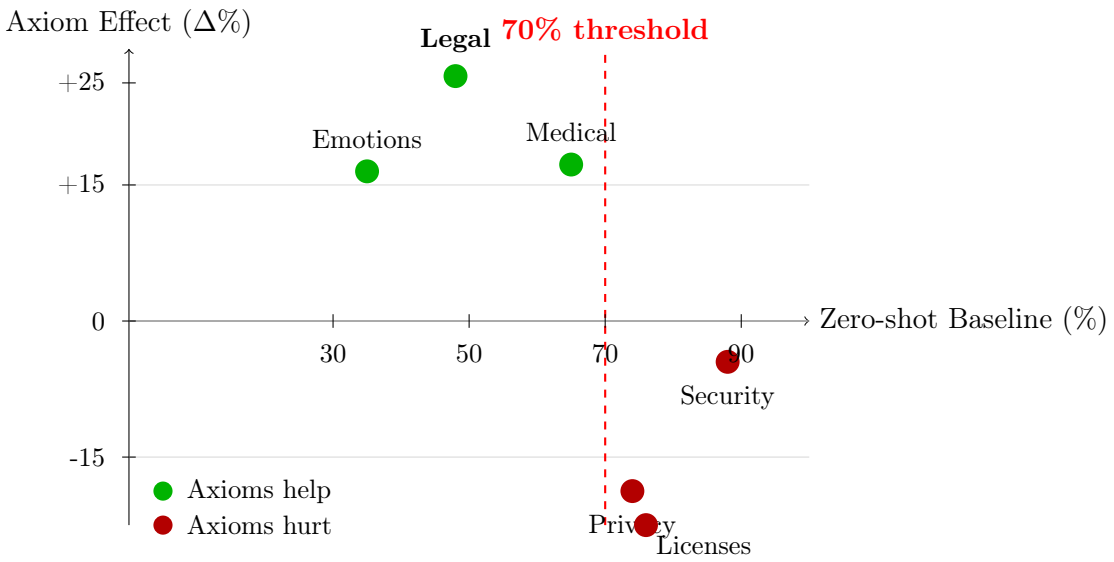

Figure 1: **The 70% Rule.** Below 70% zero-shot accuracy, axiomatic prompting improves performance (+15–24%). Above 70%, it hurts (-5 to -20%). Results from 8 models across 6 classification tasks.

The technique is simple to implement, requires no fine-tuning, and—when applied appropriately—delivers substantial accuracy gains. But applied inappropriately, it reliably degrades performance.

This paper provides the first systematic characterization of when axiomatic prompting works, validated across 8 frontier models from 5 providers.

1.1 Contributions

1. We **introduce axiomatic prompting** and provide a formal specification for axiom construction.

2. We identify the **70% threshold rule**: axioms help below this baseline, hurt above it. This rule achieved 100% prediction accuracy across 6 diverse tasks.

3. We provide **statistically significant evidence** across 8 models (GPT-4o, Claude Sonnet 4, Gemini 2.5, Llama 3.3, Qwen 2.5, and others), with all effects significant at $p < 0.05$.

4. We characterize **axiom quality factors**: pattern-based axioms outperform conceptual definitions; medium-length axioms (∼1500 tokens) outperform both shorter and longer variants.

2 Axiomatic Prompting

2.1 Definition

Definition: Axiomatic Prompting

A prompting strategy that augments classification prompts with explicit decision rules (*axioms*) mapping observable text patterns to category labels.

Form: `IF [pattern] THEN [category]`

Example: "executed in counterparts" → Counterparts

Unlike few-shot prompting (which provides examples) or chain-of-thought (which elicits reasoning), axiomatic prompting provides *rules*. The model receives explicit criteria for classification decisions.

2.2 Comparison to Related Techniques

Table 1: Prompting Strategy Comparison

Strategy	Provides	Requires	Best For
Zero-shot	Task description	Nothing	High-capability tasks
Few-shot	Examples	Labeled data	Most tasks
Chain-of-thought	Reasoning scaffold	Nothing	Complex reasoning
Axiomatic	**Decision rules**	**Domain expertise**	**Low-baseline tasks**

2.3 Axiom Construction

Effective axioms share three properties:

1. **Pattern-based**: Use observable lexical patterns ("contains X"), not abstract descriptions ("discusses the concept of")
2. **Derived from data**: Extract patterns from actual classification examples, not theoretical definitions
3. **Complete coverage**: Provide at least one axiom per category

Good axiom (pattern-based):

```
IF text contains "indemnify" OR "hold harmless"
THEN classify as "Indemnification"
```

Bad axiom (conceptual):

```
IF text describes protection from liability
THEN classify as "Indemnification"
```

The difference matters: pattern-based axioms achieved +12.4% average improvement in our experiments, while conceptual axioms averaged -16.7%.

3 The 70% Rule

Our central finding: axiomatic prompting effectiveness depends entirely on the model's baseline capability on the task.

3.1 Main Results

Table 2: The 70% Threshold: Axiom Effect by Task (gpt-4o-mini, n=100, 3 runs)

Task	Domain	Zero-shot	Axiomatic	Δ	p	Cohen's d
Below 70% baseline: axioms help						
GoEmotions	Emotions	35.0%	49.7%	**+14.7%**	0.002**	+12.70
LEDGAR	Legal	48.0%	72.0%	**+24.0%**	0.041*	+2.75
CADEC	Medical	65.0%	80.3%	**+15.3%**	0.009**	+6.09
Above 70% baseline: axioms hurt						
OPP-115	Privacy	73.7%	57.0%	**-16.7%**	0.014*	-4.75
SPDX	Licenses	76.3%	56.0%	**-20.3%**	0.005**	-8.08
CWE	Security	88.0%	83.7%	**-4.3%**	0.023*	-3.75

* $p < 0.05$, ** $p < 0.01$. All effects statistically significant.

The threshold rule achieves perfect prediction: all three tasks below 70% show significant improvement; all three tasks above 70% show significant degradation.

3.2 Cross-Model Validation

The threshold rule holds across all 8 models tested:

Table 3: Axiom Effect (Δ%) by Model and Task Domain

Model	Emotions	Legal	Medical	Privacy	Licenses	Security
GPT-4o	+10.3	+15.3	+10.3	-11.3	-40.0	+1.3
GPT-4o-mini	+16.0	+23.0	+16.7	-17.0	-18.3	-3.0
Claude Sonnet 4	+13.3	+21.3	+13.3	-3.7	+1.3	+4.7
Claude 3.5 Haiku	+3.7	+9.0	+4.0	+2.0	+4.0	+0.7
Gemini 2.5 Flash	+14.7	+3.3	+14.0	-31.7	-2.7	+35.7
Gemini 2.0 Flash	+15.7	+20.7	+6.7	-6.3	-16.3	+2.7
Llama 3.3 70B	+12.3	+16.7	+5.7	-5.0	-12.3	+6.0
Qwen 2.5 72B	+11.0	+20.3	+9.7	-8.0	-7.3	-2.0
Avg Zero-shot	33%	49%	63%	62%	60%	67%

Green-shaded tasks (Emotions, Legal, Medical) have low baselines; axioms consistently help.

Across 48 model-task combinations, the correlation between zero-shot baseline and axiom effect is $r = -0.46$ ($p < 0.001$): the better the model performs without axioms, the more axioms hurt.

4 Why the Threshold Effect?

4.1 The Interference Hypothesis

When a model already achieves high accuracy (>70%), its learned representations already capture the statistical patterns needed for classification. Adding explicit axioms introduces noise: the model must reconcile its internal representations with external rules that may be:

- **Redundant**: The model already knows "governed by" indicates Governing Laws
- **Incomplete**: Axioms can't capture every valid pattern
- **Conflicting**: Axiom formulations may differ from learned patterns

When the model struggles (<70%), axioms provide useful signal that training did not adequately capture. The rules fill knowledge gaps rather than creating conflicts.

4.2 Evidence from Ablation

We tested this hypothesis by varying axiom characteristics on a single task (GoEmotions):

Table 4: Ablation: Axiom Length and Format Effects

Condition	Axiom Length	Format	Accuracy	vs. Medium-Simple
short_simple	944 chars	Simple	30%	-12%
medium_simple	1,611 chars	Simple	**42%**	—
long_simple	6,547 chars	Simple	38%	-4%
medium_reasoning	1,611 chars	CoT	38%	-4%
medium_hybrid	1,611 chars	Hybrid	36%	-6%

Findings:

- **Length**: Medium-length axioms (∼1,600 chars) outperform both shorter (-12%) and longer (-4%)
- **Format**: Simple prompts outperform chain-of-thought (-4%) and hybrid (-6%) formats
- **Optimal**: Pattern-based axioms + simple output format

The "Goldilocks effect" suggests axioms must be detailed enough to provide useful signal, but concise enough to avoid attention dilution.

5 Practical Guidelines

Decision Flowchart for Practitioners

1. **Run zero-shot first** on a sample of your data
2. **If accuracy $< 70\%$**: Consider axiomatic prompting
   - Write pattern-based axioms ("keyword $\rightarrow$ label")
   - Derive patterns from actual data samples
   - Aim for $\sim$1,500 tokens of axioms
   - Use simple output format (no "think step by step")
3. **If accuracy $\geq 70\%$**: Use zero-shot or few-shot instead
   - Axioms will likely hurt performance
   - Few-shot provides examples without rule conflicts

5.1 Axiom Writing Checklist

Characteristic	Good	Bad
Pattern type	Lexical ("contains X")	Conceptual ("discusses")
Source	Data-derived	Theoretical
Coverage	All categories	Partial
Length	1,000–2,000 chars	<500 or >4,000
Conditions	Multiple per axiom	Single condition
Prompt format	Simple	Chain-of-thought

6 Experimental Details

6.1 Models

We tested 8 models from 5 providers:

- **OpenAI**: GPT-4o, GPT-4o-mini
- **Anthropic**: Claude Sonnet 4, Claude 3.5 Haiku
- **Google**: Gemini 2.5 Flash, Gemini 2.0 Flash
- **Meta**: Llama 3.3 70B (via HuggingFace)
- **Alibaba**: Qwen 2.5 72B (via HuggingFace)

6.2 Datasets

6.3 Protocol

- **Sample size**: 100 samples per task

Table 5: Dataset Summary

Dataset	Domain	Categories	N	Source
GoEmotions	Emotions	27	2,984	Demszky et al. [2020]
LEDGAR	Legal contracts	100	5,000	Tuggener et al. [2020]
CADEC	Medical ADRs	20	4,815	Karimi et al. [2015]
OPP-115	Privacy policies	10	6,381	Wilson et al. [2016]
SPDX	Software licenses	57	465	Software Package Data Exchange (SPDX) [2024]
CWE	Security vulns	25	5,317	Computer Incident Response Center Luxembourg (CIRCI

- **Runs**: 3 independent runs for variance estimation
- **Temperature**: 0.0 for reproducibility
- **Random seed**: 42 for sample selection
- **Strategies**: Zero-shot, few-shot (k=3), chain-of-thought, ReAct, skeleton-of-thought, axiomatic
- **Total experiments**: 864 (8 models $\times$ 6 tasks $\times$ 6 strategies $\times$ 3 runs)

7 Related Work

Prompting strategies. Chain-of-thought prompting [Wei et al., 2022] elicits intermediate reasoning steps. ReAct [Yao et al., 2023] interleaves reasoning with actions. These techniques scaffold output generation but do not provide domain-specific knowledge.

Knowledge augmentation. Retrieval-augmented generation (RAG) provides external knowledge at inference time. Mallen et al. [2023] showed that external knowledge can help or hurt depending on the model's existing capabilities. Our work extends this finding to classification rules specifically.

Rule-based NLP. Traditional NLP relied heavily on hand-crafted rules before neural methods dominated. Axiomatic prompting represents a hybrid: using rules to augment, not replace, learned representations.

8 Limitations

1. **Confounded design**: Each task has one axiom set, so axiom style is confounded with task. Future work should vary axiom quality within tasks.
2. **English only**: All experiments use English datasets. The 70% threshold may vary for other languages.
3. **Classification only**: We test only multi-class classification. Generalization to other tasks (generation, reasoning) is unknown.
4. **Extraction errors**: Some models (especially Claude 3.5 Haiku) exhibited high output parsing failure rates, introducing variance.

9 Conclusion

We introduced axiomatic prompting and characterized when it works. The answer is simple: **check zero-shot first**.

- Below 70% baseline: axioms help (+15–24%)
- Above 70% baseline: axioms hurt (-5 to -20%)

This 70% Rule held across 8 models from 5 providers, achieving 100% prediction accuracy on 6 diverse classification tasks.

The broader principle: **don't teach what the model already knows**. External rules help when learned representations are insufficient, but create interference when they're already adequate. Before investing in rule engineering, test whether your model needs the help.

Reproducibility

Code, data, axiom definitions, and full results available at: `https://doi.org/10.17605/OSF.IO/PCX2D`

- Random seed: 42
- Temperature: 0.0
- All API calls logged with timestamps
- Axiom specifications provided in supplementary materials

References

Computer Incident Response Center Luxembourg (CIRCL). Vulnerability-cwe-patch dataset. Hugging Face Datasets, 2024. URL `https://huggingface.co/datasets/CIRCL/vulnerability-cwe-patch`. 16,123 vulnerabilities with CWE classifications and patches, generated via VulnTrain.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, 2020.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. Cadec: A corpus of adverse drug event annotations. In *Journal of Biomedical Informatics*, volume 55, pages 73–81, 2015.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 2023.

Software Package Data Exchange (SPDX). Spdx license list. GitHub, 2024. URL `https://github.com/spdx/license-list-data`. Standardized license identifiers for open source software.

Don Tuggener, Pius von Däniken, Thomas Peber, and Mark Cieliebak. Ledgar: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1235–1241, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1330–1340, 2016.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.