Reasoning Power Outweighs Data Structure

A Surprising Lesson from the CyberSOCEval Malware Benchmark

A Priam Research Briefing

The Goal: Automating One of Security's Most Demanding Tasks

Large Language Models are being tested against the most formidable challenges in the Security Operations Center. We are targeting **Task 1: Malware Investigation** from the CyberSOCEval benchmark, a task requiring deep analytical reasoning beyond simple pattern matching. This benchmark builds on the foundational work of **Meta and CrowdStrike**, using real-world malware reports from the **CrowdStrike Falcon® Sandbox**.

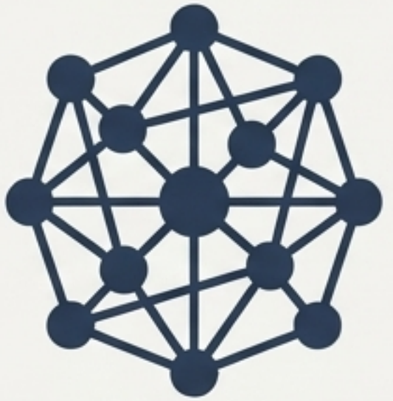
Malware Sample → Sandbox Detonation → Complex System Logs → LLM Analysis

Can Clever Formatting Unlock an LLM's Inner Detective?

A common assumption is that LLM performance hinges on data presentation. To test this, we engineered inputs using advanced, structured formats designed to enhance comprehension and reasoning.

TOON (Token-Oriented Object Notation)

A compact, schema-aware format designed to minimize tokens and improve parsing.



Knowledge Graph (KG) Encoding

Represents data as a network of entities and relationships to facilitate multi-hop reasoning, based on the 'Thinking in Graphs' methodology.

The hypothesis was simple: better-structured data should lead to better results.

A Controlled Experiment: Four Formats, One Set of Data

To isolate the impact of input structure on reasoning, we benchmarked models using four distinct representations of the exact same underlying security log data.

JSON

The default system log format.

Benefit: Leverages the inherent structure of the original data.

TOON

Token-Oriented Object Notation.

Benefit: Compact, schema-aware, improved parsing.

Graph

Knowledge Graph Encoding.

Benefit: Facilitates multi-hop reasoning.

Markdown

Standard document formatting.

Benefit: Provides a human-readable, visually organized structure.

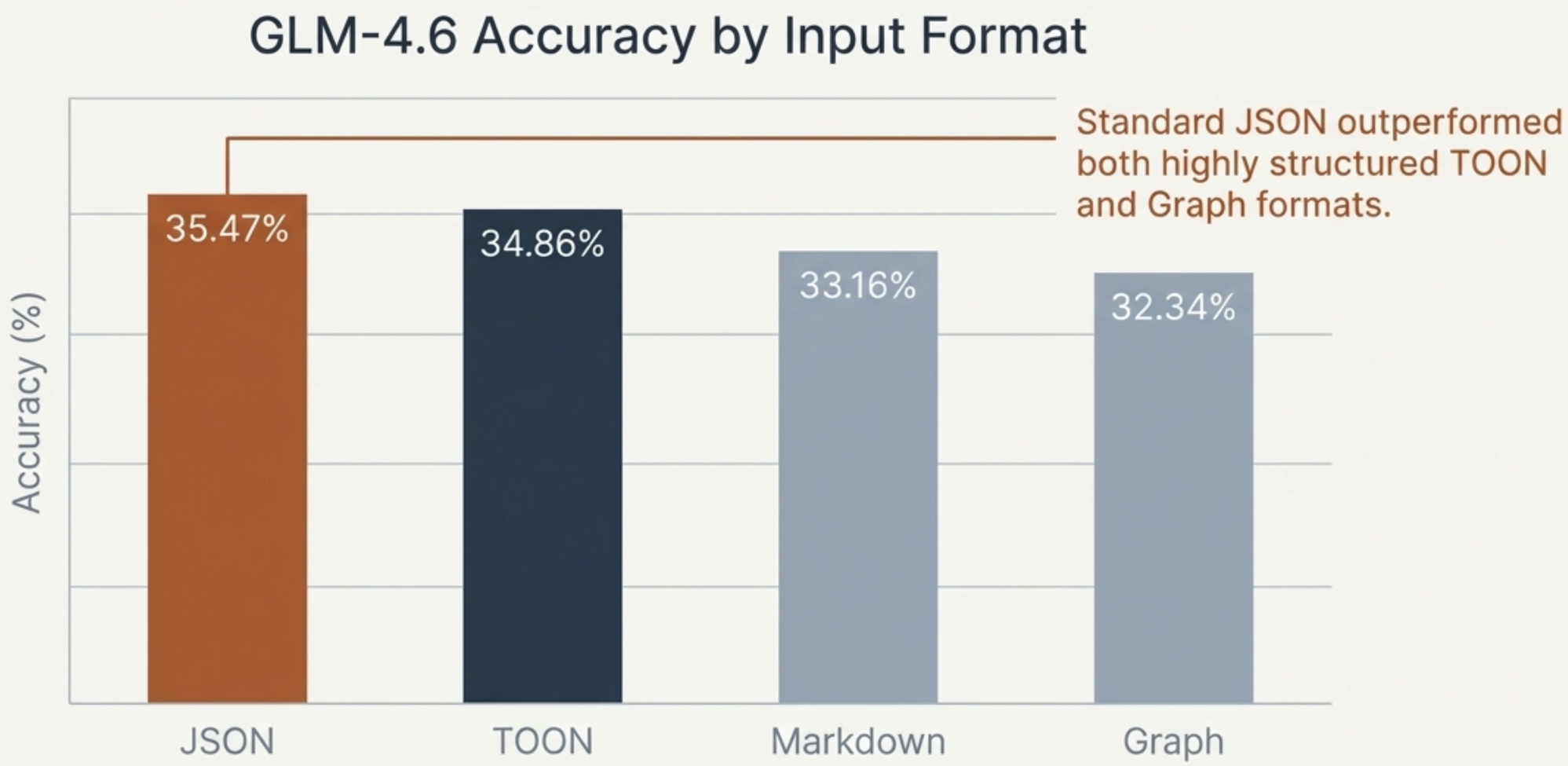
The Verdict: Advanced Formatting Offered No Advantage

Top Score: 35.47%

Achieved by **GLM-4.6** using the standard **JSON** input.

The highest accuracy was not achieved with a sophisticated format. The default, baseline representation performed best, establishing the current performance ceiling for open-source models on this task.

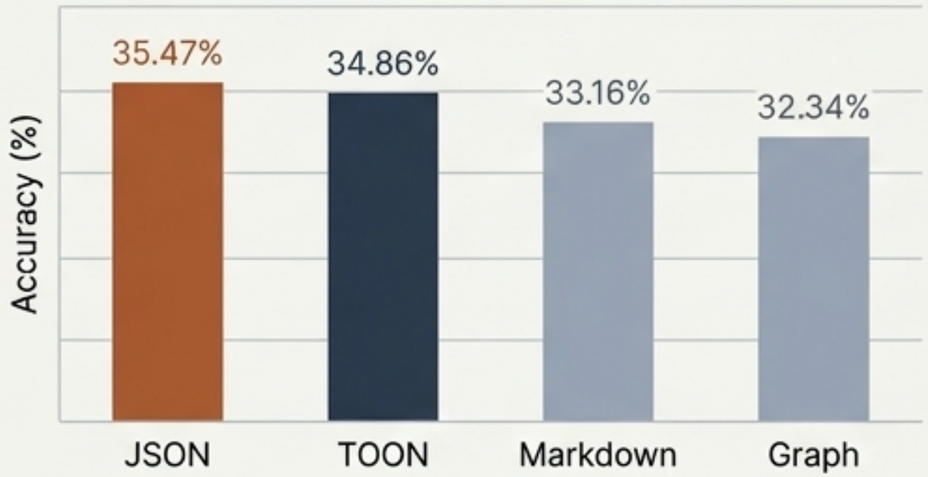
Deconstructing the Top Performer: GLM-4.6



The ‘No Free Lunch’ Principle in Data Representation

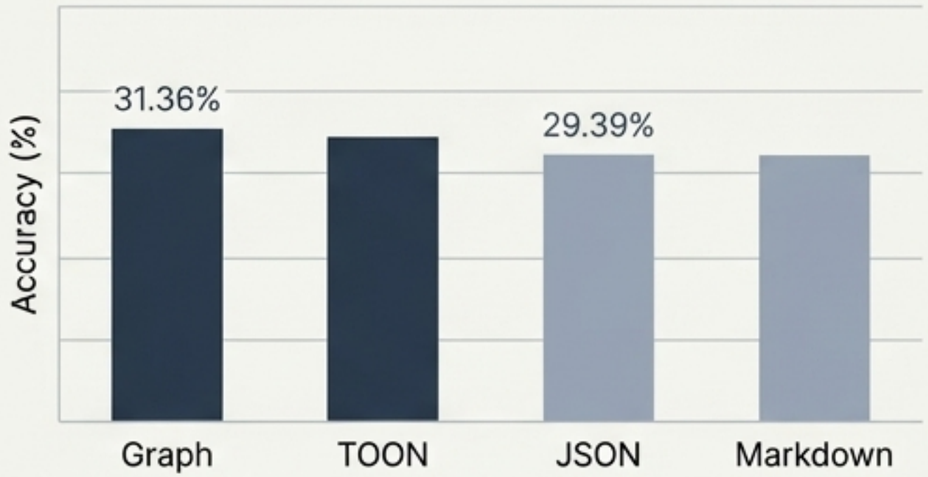
The trend holds across different model families. Structured encoding provides no consistent performance lift, suggesting the bottleneck lies elsewhere.

GLM Family



The standard **JSON** format (35.47%) slightly outperformed the highly structured TOON (34.86%) and Graph (32.34%) formats.

Claude Family



The variance was minimal. Graph encoding (31.36%) only marginally outperformed standard JSON (29.39%), with all formats clustered tightly.

Comprehensive Benchmark Results Across All Models

Model	Format	Accuracy
GLM-4.6	**JSON**	**35.47%**
GLM-4.6	TOON	34.86%
GLM-4.6	Markdown	33.16%
GLM-4.6	Graph	32.34%
Claude 3.5 Sonnet	Graph	31.36%
Claude 3.5 Sonnet	TOON	29.72%
Claude 3.5 Sonnet	JSON	29.39%
Claude 3.5 Sonnet	Markdown	27.91%
MiniMax-M2	JSON	30.00%
DeepSeek-V3.1-Terminus	JSON	28.73%
GPT-OSS 120b	JSON	28.40%
DeepSeek-V3.2-Non-reasoning	JSON	26.92%
Cisco 8b Instruct	JSON	16.42%
Cisco 8b Instruct	Graph	16.42%
Cisco 8b Instruct	Markdown	13.79%
Kimi 2 Thinking	JSON	16.09%

A New Finding Emerges: Open Source Achieves Parity

A notable outcome of the benchmark is the impressive performance of open-source models in a highly technical domain.

Open-Source Champion

GLM-4.6

35.47%

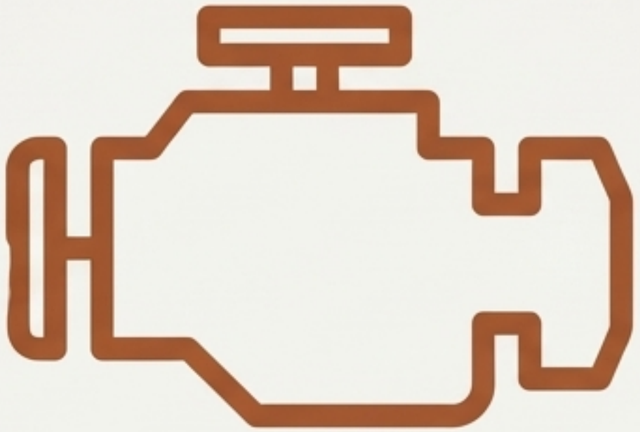
Top Proprietary Model

Claude 3.5 Sonnet

31.36%

High-quality community models like **GPT-OSS 120b** (28.40%) are also performing on par with, or exceeding, established proprietary solutions on this task.

The Bottleneck is the Engine, Not the Road



Core Reasoning Capability

Data Input Structure

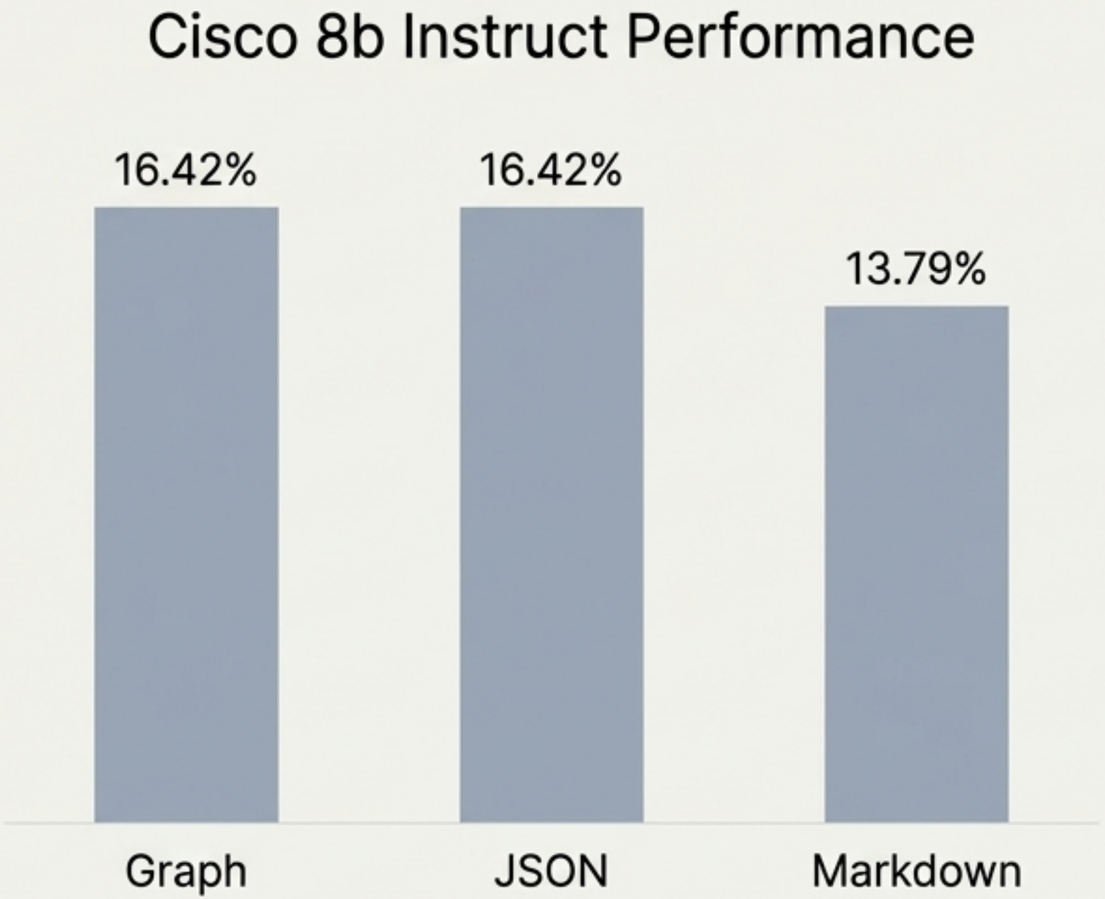
- Our findings suggest that the model’s fundamental domain reasoning capability is the primary limiting factor.
- A powerful engine can handle a basic road.
- A perfectly paved road cannot help a weak engine.

Case Study: Domain Tuning Cannot Replace Reasoning Horsepower

Focus on: The Cisco 8b Instruct model.

- This model is specifically fine-tuned for the security domain.
- Despite this specialization, its small parameter size appears to limit its raw reasoning capability.

Domain knowledge is necessary, but for complex malware investigation, it is not sufficient without the underlying reasoning horsepower.



Key Discoveries from CyberSOCEval

1. The ‘No Free Lunch’ Principle

Sophisticated data encodings like TOON and Knowledge Graphs provided no consistent performance lift. Raw reasoning outweighs clever formatting.

2. Open-Source Parity

The top-performing model, GLM-4.6, is open-source, outperforming leading proprietary models on this specific, highly technical task.

3. Reasoning is the Bottleneck

Foundational reasoning capability is more critical than domain-specific fine-tuning, especially in smaller models.

The Path Forward: Shattering the Reasoning Ceiling

With the myth of data structure debunked, our focus shifts. The true frontier for advancing automated security is not in how we present data, but in fundamentally enhancing the core reasoning capabilities of the models themselves.

We are now actively experimenting with advanced techniques to move beyond the current 35.47% ceiling and unlock the next level of performance.