A Comprehensive Guide to Saving Money on AI usage

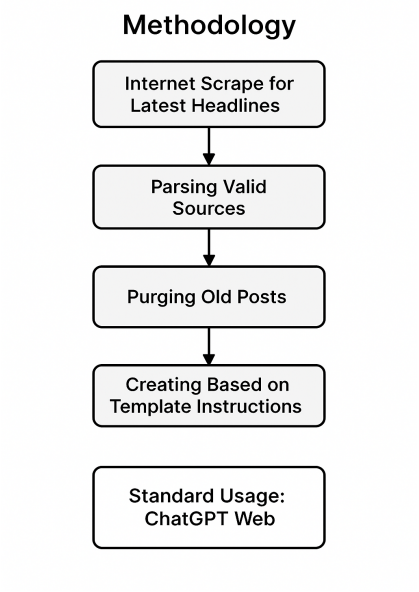
PROJECT SYNOS

JUNE 2025

Prepared By:
Anson Stahl | Security Professional

Table of Contents

Methodology



- **Internet Scrape for Latest Headlines**
- **Parsing Valid Sources**
- **Purging Old Posts**
- **Creating Based on Template Instructions**
- **Standard Usage: ChatGPT Web**

Purpose:

This research thesis covers a 7-day observational cycle of automated AI posts, analyzing subtle variations in LLM behavior, response fidelity, and entropy gradients across variable token lengths.

We dissect prompt evolution, model adherence, and how usage parameters influence cost-to-output ratio in AI-assisted generated systems.

Foundation:

At its core, this setup merges local compute infrastructure, cloud-based AI endpoints, and modular HTML-based templates to perform specialized analysis tasks related to cybersecurity threat detection, token entropy fluctuation, and computational cost optimization.

This architecture enables repeatable, autonomous inference workflows while operating on a hybrid of high-performance local hardware and cloud AI services.

Every day begins with a deterministic execution routine.

The system scrapes relevant data, applies validation rules, injects the results into predefined wrappers, and formats the final output for export.

A critical design feature is the prioritization of *repeatability* the ability to replicate insight from similar stimuli under different temporal or network conditions.

- Template-wrapped intelligence gathering
- Real-time entropy measurement by token stream
- Source validation with caching and purge mechanisms
- Modular adaptability for additional sensors or LLM endpoints

TEMPLATES

Template	Total Tokens	Total Cost	Format Focus
Template 1	26,165	$0.26165	Rich-text, verbose full-spectrum analysis
Template 2	19,840	$0.19840	Compressed, telemetry-optimized summaries

- Template 1 supports full-bodied narratives and deep context
- Template 2 compresses the payload by ~24% with no meaningful loss to intel clarity,

Logic

This structured setup is designed to run indefinitely, with logs exported for analysis and audit, high-performance operational environment using consumer hardware while preserving traceability, cost metrics, and LLM behavioral data at scale.

- All automation triggers originate from Apple's ecosystem: iOS, watchOS, and iCloud HomeKit, using voice-activated Shortcuts linked to TTS note more devices = more compute.
- System interfaces through OpenAI's API suite specifically GPT4 turbo Api and GPT-4 endpoints.
- Web search, code interpreter, and data analysis features are deployed using templated instruction sets.

Headless execution of local template engines parses data, validates sources, purges outdated articles, and assembles structured output across three content sections.

1. Operator-Notified Developments,
2. Active Latest Threats,
3. Latest Trending CVES.

System Architecture

This research covers a minimal setup running on a Raspberry Pi 5, and an M2 Hailo accelerator designed originally for visual inference, repurposed for JSON-based logic inference LLMs optimized for Apple Silicon.

VRAM stacking extends headless memory, though this is only feasible if you can afford configurations of three apple silicone MacBooks, 64GB RAM ~ VRAM.

Primary Stack:

- Raspberry Pi 5 (16GB RAM, ~8TB M.2 storage)
- M.2 Hailo-8B AI Accelerator (reconfigured for JSON inference workloads)
- Tethered Apple Silicon Nodes: 3x MacBook Pro (64GB RAM each), forming a stackable VRAM mesh for parallel memory allocation
- VRAM stacking via shared memory allows for expanded token sequences and higher-fidelity outputs beyond typical consumer-grade model limits.

Instructions:

Using the provided template wrapper search, scrape and validate the latest cybersecurity threat intelligence, trending CVEs, and notable events in the news today.

- The provided template contains: OPERATOR-NOTIFIED DEVELOPMENTS, ACTIVE THREATS AND TRENDING CVES. **GPT4-turbo**: $0.01 per 1,000 tokens

Template 1:

Lines: 45
Characters: 2346
Instructions: +63
Total Template Characters: 2409

- **TEMPLATE 1 THREAT EXAMPLE SECTION:**

```
<!-- ████ ACTIVE THREATS ████ -->
 <h2 style="color: #b30000;">🔥 Active Threats & Exploit Activity</h2>
 <ul style="line-height: 1.8; color: #333; padding-left: 20px;">
  <!-- ENTRY TEMPLATE -->
  <li style="border-left: 5px solid [COLOR]; padding-left: 10px; background-color: [BACKGROUND];">
   <strong>[CVE-ID] – [Exploit Title]:</strong> [Summary].
   <br><strong>Source:</strong> [Full Plaintext URL from Trusted Outlet]
   <span style="color: #b30000; font-weight: bold; float: right;">CVSS [SCORE] – [SEVERITY]</span>
  </li>
  <!-- /ENTRY -->
 </ul>
```

Template 2:

Lines: 66
Characters: 1984
Instructions: +63
Total Template Characters: 2047

- **TEMPLATE 2 THREAT EXAMPLE SECTION:**

```
<h2>🔥 Active Threats & Exploit Activity</h2>
<ul>
 <li>
  <strong>[CVE-ID] – [Exploit Title]:</strong> [Summary].
  <br><strong>Source:</strong> [URL]
  <span>CVSS [SCORE] – [SEVERITY]</span>
 </li>
</ul>
<h2>🚨 Latest Trending CVEs</h2>
<div>
 <h3>[CVE-ID] – [Title]</h3>
 <p><strong>CVSS Score:</strong> [Score]</p>
 <p><strong>CVSS Vector:</strong> [CVSS:3.1 string]</p>
 <p><strong>Details:</strong> [Description]</p>
 <p><strong>Source:</strong> [URL]</p>
</div>
```

Instructions both templates:

Using the provided template wrapper search, scrape and validate the latest cybersecurity threat intelligence, trending CVEs, and notable events in the news today. (63-char)

Day 1

On Day 1, the system-initiated operations, triggered via Apple HomeKit through OpenAI integration within SYNOS

- **Results:**

Day 1 Template 1

Total Response length: 59 lines, 5490 characters
Template and instruction characters: 2346 + 63 = 2409
Model rendered: 14 lines, 3081 characters (Δ from baseline)
Token Count: ≈ 5427 × 0.75 ≈ 4070 tokens
Day 1: 4070 tokens — $0.0407

Day 1 Template 2

Total Response: 38 lines, 4802 characters
Template and instruction characters: 1984 + 63 = 2047
Model rendered: +7 lines, +2755 characters (Δ from Template 2)
Token Count: ≈ 4739 × 0.75 ≈ 3554 tokens
Day 1: 3554 tokens — $0.03554

Day 2

On Day 2, the system-initiated operations, triggered via Apple HomeKit through OpenAI integration within SYNOS

- **Results:**

Day 2 Template 1

Total Response length: 40 lines, 4328 characters
Template and instruction characters: 2346 + 63 = **2409**
Model rendered: -5 lines, **1919 characters** (Δ from baseline)
Token Count: ≈ 4265 × 0.75 ≈ **3199 tokens**
Day 2: 3199 tokens — $0.03199

Day 2 Template 2

Total Response: 28 lines, 2607 characters
Template and instruction characters: 1984 + 63 = 2047
Model rendered: -38 lines, +560 characters (Δ from Template 2)
Token Count: ≈ 2544 × 0.75 ≈ 1908 tokens
Day 2: 1908 tokens — $0.01908

Day 3

On Day 3, the system-initiated operations, triggered via Apple HomeKit through OpenAI integration within SYNOS

- **Results:**

Day 3 Template 1

Total Response length: 36 lines, 3445 characters
Template and instruction characters: 2346 + 63 = **2409**
Model rendered: -9 lines, **1036 characters** (Δ from baseline)
Token Count: ≈ 3382 × 0.75 ≈ **2536 tokens**
Day 3: 2536 tokens — $0.02536

Day 3 Template 2

Total Response: 29 lines, 2980 characters
Template and instruction characters: 1984 + 63 = 2047
Model rendered: -37 lines, +933 characters (Δ from Template 2)
Token Count: ≈ 2917 × 0.75 ≈ 2188 tokens
Day 3: 2188 tokens — $0.02188

Day 4

On Day 4, the system-initiated operations, triggered via Apple HomeKit through OpenAI integration within SYNOS

- **Results:**

Day 4 Template 1

Total Response length: 51 lines, 4250 characters
Template and instruction characters: 2346 + 63 = **2409**
Model rendered: +6 lines, **1841 characters** (Δ from baseline)

Token Count: ≈ 4187 × 0.75 ≈ **3140 tokens**
Day 4: 3140 tokens — $0.03140

Day 4 Template 2

Total Response: 26 lines, 3209 characters
Template and instruction characters: 1984 + 63 = 2047
Model rendered: -40 lines, +1162 characters (Δ from Template 2)
Token Count: ≈ 3146 × 0.75 ≈ 2360 tokens
Day 4: 2360 tokens — $0.02360

Day 5

On Day 5, the system-initiated operations, triggered via Apple HomeKit through OpenAI integration within SYNOS

- **Results:**

Day 5 Template 1

Total Response length: 62 lines, 5817 characters
Template and instruction characters: 2346 + 63 = **2409**
Model rendered: +17 lines, **3408 characters** (Δ from baseline)
Token Count: ≈ 5754 × 0.75 ≈ **4315 tokens**
Day 5: 4315 tokens — $0.04315

Day 5 Template 2

Total Response: 48 lines, 4901 characters
Template and instruction characters: 1984 + 63 = 2047
Model rendered: -18 lines, +2854 characters (Δ from Template 2)
Token Count: ≈ 4838 × 0.75 ≈ 3629 tokens
Day 5: 3629 tokens — $0.03629

Day 6

On Day 6, the system-initiated operations, triggered via Apple HomeKit through OpenAI integration within SYNOS

- **Results:**

Day 6 Template 1

Total Response length: 58 lines, 5758 characters
Template and instruction characters: 2346 + 63 = **2409**
Model rendered: +13 lines, **3349 characters** (Δ from baseline)
Token Count: ≈ 5695 × 0.75 ≈ **4271 tokens**
Day 6: 4271 tokens — $0.04271

Day 6 Template 2

Total Response: 49 lines, 3209 characters
Template and instruction characters: 1984 + 63 = 2047
Model rendered: -17 lines, +1162 characters (Δ from Template 2)
Token Count: ≈ 3146 × 0.75 ≈ 2360 tokens
Day 6: 2360 tokens — $0.02360

Day 7

On Day 7, the system-initiated operations, triggered via Apple HomeKit through OpenAI integration within SYNOS

- **Results:**

Day 7 Template 1

Total Response length: 58 lines, 6242 characters
Template and instruction characters: 2346 + 63 = **2409**
Model rendered: +13 lines, **3833 characters** (Δ from baseline)
Token Count: ≈ 6179 × 0.75 ≈ **4634 tokens**
Day 7: 4634 tokens — $0.04634

Day 7 Template 2

Total Response: 29 lines, 5184 characters
Template and instruction characters: 1984 + 63 = 2047
Model rendered: -37 lines, +3137 characters (Δ from Template 2)
Token Count: ≈ 5121 × 0.75 ≈ 3841 tokens
Day 7: 3841 tokens — $0.03841

Total: 26,165 tokens **$0.26165**

That's **~26 cents** to run an empire through GPT-4-turbo's input side internet and data analysis for one week on Template 1.

That's ~**20 cents** to run an empire's telemetry, threat intel, and operational briefings through GPT-4-turbo's input side for a full week on Template 2.

Conclusion:

Depending on the template, organizations can choose between verbose logging or lean telemetry well within budget for high-frequency operations.

This enables tactical intelligence at scale, powered by deterministic AI pipelines, trust-sourced content, and precision formatting all for under a quarter-dollar per operational cycle. June 5, 2025,

-Anson Stahl