Website Data Collection Optimization Report

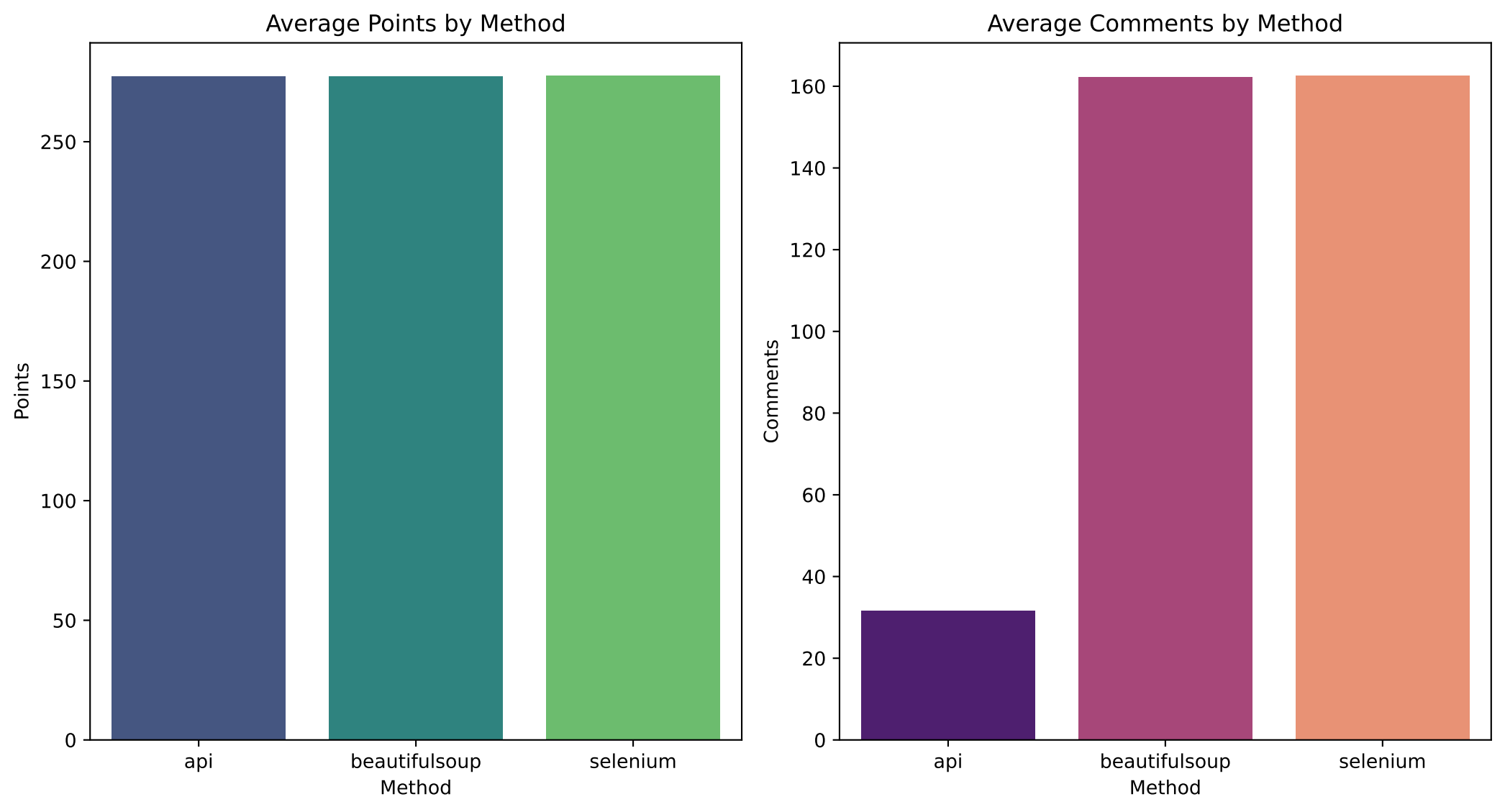
Target Site: https://news.ycombinator.com
Data Collected: Titles, points, comment counts, authors, first comment

Key Highlights
• Total requests issued: 190
• Aggregate bandwidth consumed: 5584.3 KB
• Fastest method: Api
• Selenium captured full rendered context (comment text) at the cost of higher latency.

Objective:
Compare scraping efficiency, observe network behaviour, and recommend the most resilient workflow.

Performance & Network Summary

method	total_time_s	total_requests	total_bytes	avg_latency_ms
beautifulsoup	32.60	21	4833.1 KB	668.5
api	5.12	41	32.5 KB	295.2
selenium	49.31	128	718.6 KB	112.9


Average Points by Method
Points
0
50
100
150
200
250
api
beautifulsoup
selenium
Method
Average Comments by Method
Comments
0
20
40
60
80
100
120
140
160
api
beautifulsoup
selenium
Method

Headline Keyword Signals

```
Positive Indicators (more likely to rank high):
aws                   (0.80)
code                  (0.75)
code web              (0.75)
claude code           (0.75)
claude                (0.75)
video                 (0.52)
fps video             (0.52)
fps                   (0.52)
pointer               (0.52)
pointer 2b            (0.52)

Negative Indicators (less likely to rank high):
hn                    (-0.32)
practical             (-0.30)
practical scheme      (-0.30)
scheme                (-0.30)
compiler              (-0.27)
led panel             (-0.23)
panel                 (-0.23)
led                   (-0.23)
small                 (-0.23)
small led             (-0.23)
```

Recommended Strategy & Hardening Checklist

Optimal Workflow

- Use the API collector for frequent polling (fastest: Api).
- Augment with the BeautifulSoup scraper to capture rendered comment context.
- Schedule Selenium runs hourly to validate UI changes and keep parsing selectors fresh.

Hardening Steps

- Enforce rate limiting via the configurable `throttle_s` arguments.
- Restrict outbound ports with `ufw` during tests to ensure graceful degradation.
- Capture traffic with `tcpdump` and archive `.pcap` files in `network/` for audits.
- Route high-volume runs through a proxy or VPN and refresh credentials securely.