API Rate Limiting & Throttling Specification

Overview

This document describes the rules and expected behavior around rate limiting and throttling for consuming the Example API.

Definitions

- **Rate Limiting**: Restricting the number of API requests a client can make in a given timeframe.
- **Throttling**: Delaying or denying API requests once rate limits are exceeded.

Limits

Plan	Limit	Time Window	Scope
Free	60 requests	per minute	per API key
Pro	600 requests	per minute	per API key
Enterprise	Custom	per minute	per API key

Headers

Header	Description
`X-RateLimit-Limit`	Maximum number of requests allowed in the current time window.
`X-RateLimit-Remaining`	Remaining requests in the current window.
`X-RateLimit-Reset`	Timestamp (UTC) when the rate limit resets.

Behavior

- Once the limit is reached, subsequent requests receive `HTTP 429 Too Many Requests`.
- Clients should use `X-RateLimit-Reset` to determine when to retry.
- No partial credits are carried over to the next window.

Sample Response

```
HTTP/1.1 429 Too Many Requests
Content-Type: application/json
X-RateLimit-Limit: 60
X-RateLimit-Remaining: 0
X-RateLimit-Reset: 1719981600
{
  "error": "Rate limit exceeded. Try again later."
}
```

Best Practices

- Use appropriate API keys for each application instance.
- Monitor returned rate limit headers and back off accordingly.
- If you require higher limits, contact support for an upgrade.