Claude Rate Limits: Every Plan & API Tier (2026)

Edition of November 13, 2026 — facts verified November 13, 2026. Current edition always at https://tygartmedia.com/claude-rate-limits/.

Direct Answer (November 13, 2026): Claude rate limits scale by plan: Free accounts get dynamic session limits; Pro ($20/mo) provides at least 5x the Free plan's capacity per 5-hour session (Anthropic publishes no fixed message counts; usage varies with message length, attachments, conversation length, and model); Max ($100/mo) and Ultra ($200/mo) offer extended quotas for Claude Code. API rate limits operate on rolling token buckets under the current spend-based usage tiers (Evaluation, Start, Build, Scale, and Custom), with cached tokens exempt from input rate caps.

Updated November 2, 2026 — added direct answers on Claude Pro usage limits and reset windows, sourced from Anthropic's published pricing.

Updated October 28, 2026: Made infographic diagrams clickable to open full-size images.

Last verified: November 13, 2026 (Pacific Time). November 2026 update: Anthropic has replaced the deposit-based API tier system — the corrections are noted in the sections below, and two older tables are marked as the retired system's record.

Official links: Live status page · API docs · Upgrade plans · Help center

November 2026 note: Anthropic replaced its legacy deposit-based Tier 1–4 API limits with spend-based usage tiers (Evaluation, Start, Build, Scale, Custom), and **Claude Fable 5.2** ($8/$40 MTok) now leads the model lineup. The tables below reflect current published API rate limits.

Claude AI · Fitted Claude

Claude rate limits are the single most complained-about aspect of the product. A viral Reddit post on the topic received over 1,060 upvotes. This guide explains what the limits are at every plan tier, why they exist, and every community-tested strategy for getting more out of your plan before hitting the wall.

1. Anthropic API Usage Tier Qualifications & Thresholds

Tiers climb with spend and reliability — confirm live console limits.

Your API account's rate limits are spend-based. The current system has five tiers — Evaluation, Start, Build, Scale, and Custom — and your tier follows your monthly spend level rather than a deposit balance:

Usage Tier	Monthly Spend Cap	Notes
Evaluation	May start as low as $5 or $500	Starting tier; the initial cap depends on risk checks
Start	$500	
Build	$1,000	
Scale	$200,000	
Custom	No predefined cap	Arranged with Anthropic

Per-model RPM and TPM limits for each tier are published in the Console and in Anthropic's rate-limits documentation. This table replaces the retired deposit-based Tier 1–4 table ($5 deposit, cumulative-spend thresholds) in November 2026.

2. Requests Per Minute (RPM) and Tokens Per Minute (TPM) by Model

RPM/TPM differ by model — shapes matter more than memorized tables.

Rate limits apply independently across model families. High-intelligence models (Opus) have tighter token concurrency caps than lightweight models (Haiku):

Anthropic publishes standard per-minute limits for the current usage tiers, read from Anthropic's rate-limits documentation in October 2026. Limits are requests per minute (RPM), input tokens per minute (ITPM), and output tokens per minute

(OTPM), enforced per model class at the organization level:

Model	Start	Build	Scale
Claude Opus 5.5	1,000 / 2,000,000 / 400,000	5,000 / 5,000,000 / 1,000,000	10,000 / 10,000,000 / 2,000,000
Claude Sonnet 5.5	1,000 / 2,000,000 / 400,000	5,000 / 5,000,000 / 1,000,000	10,000 / 10,000,000 / 2,000,000
Claude Haiku 5.5	1,000 / 2,000,000 / 400,000	5,000 / 5,000,000 / 1,000,000	10,000 / 10,000,000 / 2,000,000
Claude Fable 5.2 (and prior Fable 5.1 / 5 combined)	1,000 / 500,000 / 100,000	2,000 / 1,500,000 / 300,000	4,000 / 4,000,000 / 800,000

Models still available from the prior generation (Sonnet 5, Opus 5, Haiku 4.5) carry the same standard Start limits. New organizations may begin in an Evaluation tier with lower limits that Anthropic does not publish in the standard table, and the Custom tier's limits are arranged with Anthropic. Your Console Rate limits page shows your account's actual tier and limits.

3. Diagnosing and Handling HTTP 429 Rate Limit Errors

429 is a pause — backoff, then retry with smaller batches.

When your application exceeds either its Requests-Per-Minute or Tokens-Per-Minute cap, the Anthropic API responds with an `HTTP 429 Too Many Requests` error containing response headers detailing when capacity will reset:

- `retry-after`: Number of seconds to wait before retrying.
- `anthropic-ratelimit-requests-remaining`: Remaining requests available in the current 60-second window.
- `anthropic-ratelimit-tokens-remaining`: Remaining token budget available in the current window.
- `anthropic-ratelimit-tokens-reset`: ISO timestamp indicating when the token pool will fully refresh.

Production Rate Limit Mitigation Playbook

1. **Exponential Backoff with Full Jitter:** Never retry immediately in a tight loop. Implement an exponential backoff formula with randomized jitter to prevent thundering herd spikes on your backend.
2. **Utilize Prompt Caching:** Cached prefix tokens read from memory bypass standard token generation latency and dramatically streamline token processing windows. Read our full Claude AI Pricing and Token Rates Guide for complete caching cost structures.
3. **Route Heavy Jobs to the Batch API:** For bulk processing, offline report generation, and data extraction, use the Anthropic Messages Batch endpoint. Batch jobs run against separate capacity pools, avoiding live interactive rate caps while cutting token costs by 50%.

Why Rate Limits Exist

Claude's rate limits are primarily about compute capacity, not money. Running frontier reasoning models like Claude Fable 5.2 and Opus 5.5 on complex tasks requires heavy GPU compute. Anthropic limits usage to ensure consistent performance for all users. The limits are enforced per rolling time window, not per calendar day.

Rate Limits by Plan

Free Plan

Dynamic session limits subject to platform load (upgrade to paid plans for guaranteed capacity and Claude Code). Heavy users hit limits after 5-10 substantive prompts. Anthropic adjusts dynamically based on system load.

Claude Pro ($20/month)

Roughly 5x the usage of free. Community consensus: approximately **12 heavy prompts per session before throttling**. Light prompts run much longer before hitting limits.

Claude Max 5x ($100/month)

Approximately 5x Pro limit. Anthropic publishes no per-window token figures; paid plans carry a five-hour session limit and a weekly usage limit across all models.

Claude Ultra ($200/month)

10x the Pro weekly limit, with no 5-hour limit for verified accounts. The top tier for developers running Claude Code for extended sessions and professionals processing large document volumes daily.

API Rate Limits (Current Usage Tiers)

API limits are measured in requests per minute (RPM), input tokens per minute (ITPM), and output tokens per minute (OTPM), enforced per model class at the organization level. The current usage tiers and their monthly spend caps are in section 1 above; the standard per-minute limits are in section 2. Your tier advances automatically with usage history and account standing — not by deposit or cumulative credit purchases; the deposit-based Tier 1–4 system is retired. Earlier versions of this page reproduced the retired system's tables in this section; they were removed in the November 5, 2026 rebuild because the system they describe no longer exists, and the retired figures remain available in this page's revision history.

Cache-aware ITPM: for current models, only *uncached* input tokens count toward your ITPM limit — `cache_read_input_tokens` do not. With an 80% cache-hit rate against a 2,000,000 ITPM limit you can effectively process ~10,000,000 total input tokens per minute, so prompt caching is the single best lever for raising effective throughput.

When you hit a limit, the API returns a `429` with a `retry-after` header (seconds to wait), plus `anthropic-ratelimit-*` headers showing remaining requests/tokens and reset times. Limits use a token-bucket algorithm — capacity replenishes continuously rather than resetting at a fixed clock time. The Message Batches API and Managed Agents endpoints have their own separate limits.

When are Claude's usage limits on the Pro plan?

Claude Pro usage limits operate under a two-tiered allowance structure: a short-term rolling session window that resets every 5 hours, and an overarching weekly cap that resets once per week at a fixed schedule specific to your account:

- **5-Hour Rolling Session Window:** Pro plan users receive at least 5 times the usage capacity of the free tier per session. Your 5-hour limit timer begins when you send your first message in an active session block, and full session capacity replenishes continuously 5 hours later.
- **Weekly Account Cap:** In addition to the 5-hour session allowance, Pro accounts are subject to a weekly usage limit that governs total consumption across all models, Projects, and tools (including Claude Code and integrated desktop workflows — Cowork is now part of Claude). This limit resets at a fixed, recurring day and time assigned to your account, which you can monitor under Settings > Usage.
- **Factors That Consume Limits Fastest:** Anthropic's documentation highlights three primary drivers of rapid usage consumption: *Model choice:* Flagship reasoning models like Claude Opus consume your session budget substantially faster than everyday models like Sonnet or Haiku. *Conversation length:* Because Claude re-processes the entire conversation history with every new message, long extended threads consume progressively larger token volumes on each subsequent turn. *File attachments:* Attaching large codebases, PDFs, or data files significantly increases the prompt token size sent on every turn within that chat.

Where Anthropic does not publish fixed message counts, capacity scales dynamically based on real-time platform traffic. When limits are reached, users can either wait for the rolling reset or upgrade to Claude Max for 5x expanded session and weekly headroom.

Community-Tested Workarounds

- **Use Projects with persistent system prompts** — reduces token overhead per conversation
- **Use Sonnet 5.5 for routine tasks, Opus 5.5 for complex architecture, and Fable 5.2 for the most demanding flagship work** — don't burn your limit budget on tasks Sonnet handles equally well
- **Batch related work into single long sessions** — starting five conversations uses more overhead than one long one
- **Compress your inputs** — extract only relevant sections from long documents before pasting
- **Use the API for high-volume predictable workflows** — more limit-efficient than the consumer interface for automated tasks

Frequently Asked Questions

How many messages can I send on Claude Pro?

No published exact number — depends on message complexity. Community estimates suggest roughly 12 heavy messages per session before throttling begins on Pro.

Do Claude rate limits reset daily?

Rate limits use a rolling time window, not a fixed midnight reset.

Primary & Reference Links Cited on This Page

- Live status page — `https://status.anthropic.com`
- API docs — `https://platform.claude.com/docs`
- Upgrade plans — `https://claude.com/pricing`
- Help center — `https://support.claude.com`
- Tygart Media: https://tygartmedia.com/wp-content/uploads/2026/08/claude-api-usage-tier-ladder.jpg
- Tygart Media: https://tygartmedia.com/wp-content/uploads/2026/08/claude-rate-limits-capacity-by-seat.jpg
- Tygart Media: https://tygartmedia.com/wp-content/uploads/2026/08/claude-rate-limits-429-pause.jpg
- Tygart Media: Claude AI Pricing and Token Rates Guide
- Tygart Media: https://tygartmedia.com/wp-content/uploads/2026/08/claude-rate-limits-why-ceiling.jpg