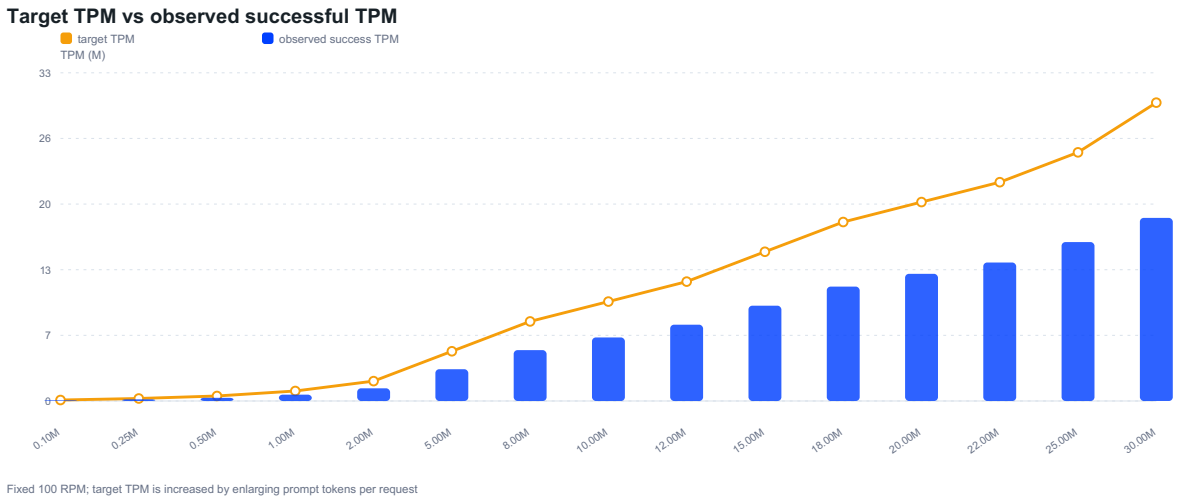
inferx Fixed 100 RPM TPM Ramp Test Report

Executive Summary

- Test conclusion: under the same fixed `100 RPM`, `0.10M -> 30.00M` target TPM ramp setup, `inferx` reached a highest zero-error target of `25.00M` and a highest observed success TPM of `18.41M` at `30.00M` target TPM.
- Stability boundary: the first degraded stage was `22.00M target TPM`; the overall success rate was `99.67%`, with a Wilson 95% CI of `99.22% – 99.86%`; total requests / success / failed were `1500 / 1495 / 5`.
- Issues encountered: this run observed 429 responses; failures were mainly `non-2xx streaming response`, and the first timeout or network-error stage was `none`.
- Latency risk: the highest p95 TTFT was `7.35s` at `30.00M target TPM`; high-TPM stages showed substantial latency and error-rate volatility, so observed success TPM should not be read alone.
- Cache reading: Successful responses returned `106,788,608` cache-read tokens in total, for an overall cache read ratio of `99.81%`.
- Ramp shape: The curve was non-monotonic: lower-TPM stages showed 429 responses, while later 25.00M stages completed successfully. Treat this run as a current rate-limit/window and queueing observation rather than a standalone absolute TPM ceiling.



TPM ramp overview

Research Questions

- Under a fixed RPM, does increasing per-request prompt size produce repeatable throughput degradation or 429 rate limiting on `inferx`?
- How large is the admission gap between `attempted target TPM` and `observed success TPM`?
- Is degradation monotonic with target TPM, or does the provider show non-monotonic rate-limit/window or queueing behavior?

Methodology

Controlled variable	Setting
Endpoint	`/v1/chat/completions` streaming
Model	`xiaomi/mimo-v2.5`
Provider lock	`provider.only=["inferx"]`, `allow_fallbacks=false`, `require_parameters=true`
Ramp mode	continuous ramp until clear degradation
RPM control	fixed `100 RPM`
Target TPM ladder	`0.10M -> 30.00M`
Stage duration	`60` seconds
Stage cooldown	`60` seconds
Network	direct `httpx trust_env=false`

Response variables include request admission rate, HTTP 429 rate, timeout rate, observed success TPM, token admission ratio, and p50/p95 TTFT plus end-to-end latency for successful requests. Success and error rates use Wilson score 95% confidence intervals to reduce over-interpretation of small per-stage samples.

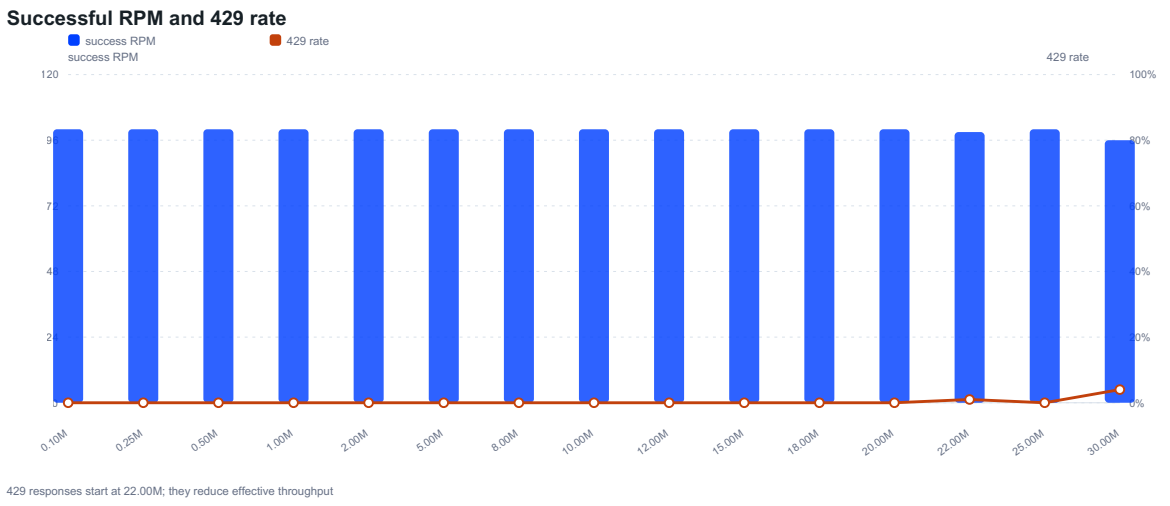
Summary

Finding	Result
Highest target TPM tested	30.00M

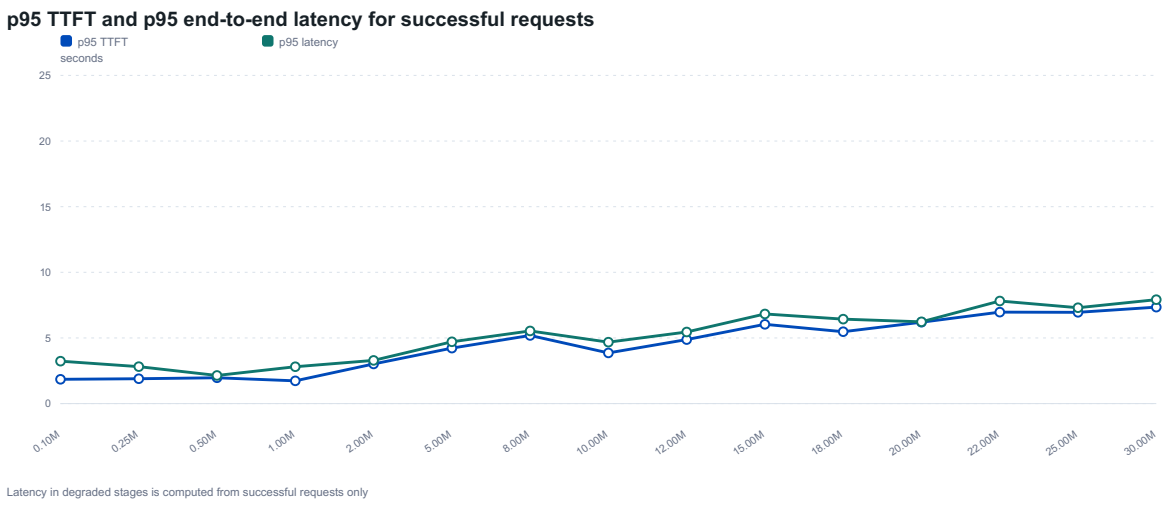
Highest zero-error target TPM	25.00M
Highest observed success TPM	18.41M @ 30.00M target
First degraded stage	22.00M target TPM
Total requests / success / failed	1500 / 1495 / 5
Overall success rate Wilson 95% CI	99.22% - 99.86%
Overall error rate Wilson 95% CI	0.14% - 0.78%
Overall cache read ratio	99.81%
Model / Provider	`xiaomi/mimo-v2.5` / `inferx`
API / network	`/v1/chat/completions` streaming / direct `trust_env=false`

This run fixed request rate at `100 RPM` and increased per-request prompt size to ramp from `0.10M -> 30.00M` target TPM. Failures first appeared at `22.00M` target TPM; the highest zero-error target was `25.00M`; the highest observed success TPM was `18.41M`. Under the current account and provider limit state, the stable throughput boundary is still affected by large-packet concurrent requests.

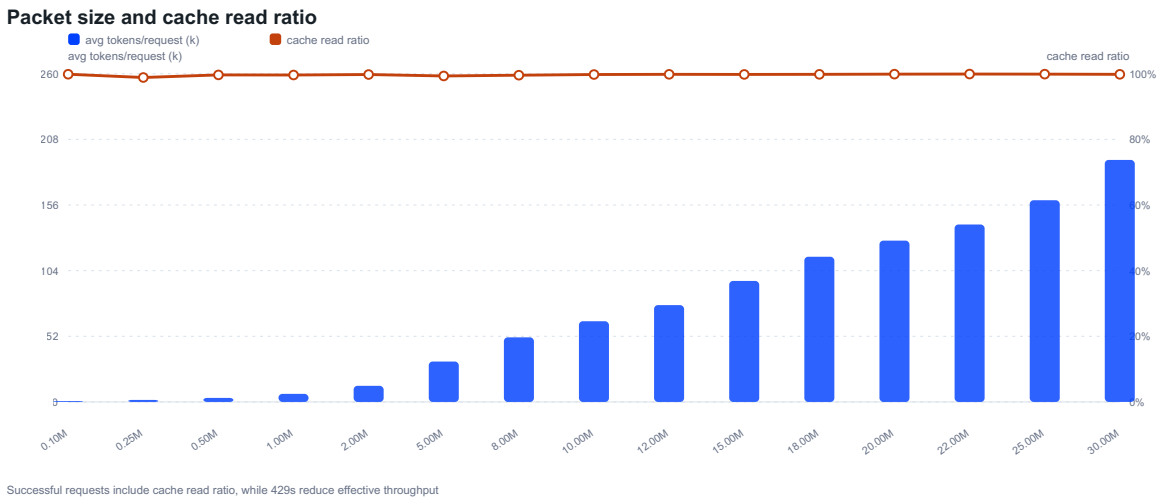
Charts



Success RPM and 429 rate



TTFT and latency for successful requests



Cache hit and packet size

Stage Results

target TPM	target RPM	success RPM	success	success 95% CI	429 rate	observed success TPM	token admission	avg toke req
0.10M	100	100	100/100	96.30%-100.00%	0.00%	0.07M	65.70%	657
0.25M	100	100	100/100	96.30%-100.00%	0.00%	0.16M	64.08%	1,60
0.50M	100	100	100/100	96.30%-100.00%	0.00%	0.32M	63.90%	3,19
1.00M	100	100	100/100	96.30%-100.00%	0.00%	0.64M	64.08%	6,40

2.00M	100	100	100/100	96.30%-100.00%	0.00%	1.28M	63.90%	12,7
5.00M	100	100	100/100	96.30%-100.00%	0.00%	3.20M	63.95%	31,9
8.00M	100	100	100/100	96.30%-100.00%	0.00%	5.11M	63.93%	51,1
10.00M	100	100	100/100	96.30%-100.00%	0.00%	6.39M	63.92%	63,9
12.00M	100	100	100/100	96.30%-100.00%	0.00%	7.67M	63.91%	76,6
15.00M	100	100	100/100	96.30%-100.00%	0.00%	9.59M	63.91%	95,8
18.00M	100	100	100/100	96.30%-100.00%	0.00%	11.50M	63.91%	115,(
20.00M	100	100	100/100	96.30%-100.00%	0.00%	12.78M	63.91%	127,8
22.00M	100	99	99/100	94.55%-99.82%	1.00%	13.92M	63.27%	140,5
25.00M	100	100	100/100	96.30%-100.00%	0.00%	15.98M	63.91%	159,7
30.00M	100	96	96/100	90.16%-98.43%	4.00%	18.41M	61.36%	191,7

Notes

- `observed success TPM` counts only successful responses with returned `usage.total_tokens`; 429 requests are excluded.
- Failed requests should be interpreted using the 429 rate, timeout rate, and sample errors.
- 429 responses first appeared at `22.00M` target TPM.
- Each stage in this run had at least 5 successful samples, so the latency percentiles are directionally readable.
- Provider was locked with `provider.only=["inferx"]`, `allow_fallbacks=false`, `require_parameters=true`, and direct networking via `httpx trust_env=false`.
- Ramp mode: continuous ramp until clear degradation; stop reason: reached max_auto_target_tpm=30000000.
- `token admission` = `observed success TPM / target TPM`; it measures the share of attempted token throughput that was admitted and completed successfully.

Threats To Validity

- Short per-stage durations make results sensitive to 429 windows, provider scheduling, and account-level transient state; a single run does not prove a durable capacity ceiling.
- Target TPM is generated from prompt-token estimates and fixed RPM; actual throughput should be read from successful response `usage.total_tokens`.
- TTFT and latency percentiles are conditional on successful requests only; they do not directly represent the user experience of failed requests.
- Failed responses may have `unknown` provider attribution. The provider lock proves routing intent, but failure attribution should still be cross-checked with response bodies and gateway logs.

Reproducibility

Item	Path
Run directory	`export/infron_provider_tpm_ramp/xiaomi_mimo_v2_5_inferx_dynamic_rpm100_tpm_ramp_30m_20260723`
Experiment manifest	`export/infron_provider_tpm_ramp/xiaomi_mimo_v2_5_inferx_dynamic_rpm100_tpm_ramp_30m_20260723/manifest.json`
Load-test script	`scripts/loadtest_infron_provider_tpm_ramp.py`
Orchestration script	`scripts/run_baidu_cloud_fixed_rpm100_tpm_ramp.py`
Report renderer	`scripts/render_fixed_rpm100_tpm_ramp_report.py`