FEAT Search API – Performance & Volumetrics Plan (Private Beta)

Scope

This round of performance testing focuses on the **FEAT Search API**.

We are not testing UI rendering performance here. The goal is to understand how the **API behaves under load**, particularly around search, filtering, pagination, details, and location/autocomplete endpoints.

This is about backend behaviour and reliability ahead of private beta.

Why we're doing this

The purpose of this testing is simple:

To understand how the FEAT Search API behaves under realistic private beta usage — and to know where the limits are before we open access more widely.

Specifically, we want to:

- Establish a clear performance baseline
- See how the API behaves under expected peak usage
- Understand how it degrades under higher-than-expected traffic
- Identify any bottlenecks early
- Make informed decisions about onboarding and monitoring

This is not about chasing perfect numbers.
It's about knowing what happens under pressure — and being prepared.

Where the traffic numbers come from

We don't yet have real production traffic data for FEAT. So the volumetrics are based on product assumptions for private beta.

These are planning assumptions only and will be refined once real usage data becomes available.

Private beta cohort

- Approx. **250 users** have been invited.
- Access is controlled.
- There is no public campaign at this stage.
- Other users may be brought on as the private beta progresses

We assume:

- 30–50% of invited users may be active on a typical day.
- A typical engagement pattern for this type of roll out is to see usage peak in the few days after the link is provided and then stabilise before declining as the more interested parties have been active.
- That gives us roughly **75–125 active users per day**.
- Each active user performs around **5–6 searches per session**.

That results in approximately:

- **375–750 searches per day**.

Assuming usage is concentrated across typical working and waking hours (around 12 hours), that works out to:

- **30–60 searches per hour** during quieter periods
- Roughly **0.01–0.02 requests per second**

Expected peak usage

We also assume that during the busiest hour:

- 10–20% of invited users may be active at once.
- That's around **25–50 users**.
- At 5–6 searches per user, this results in:

125–300 searches per hour,
which equates to approximately **0.04–0.08 requests per second sustained**.

This is our expected upper bound for normal private beta traffic.

Load levels we will test

To make testing meaningful, we will test across several bands.

1. Baseline (expected quiet usage)

- ~0.01–0.02 rps
- Used to establish normal response times and system behaviour

2. Expected peak (busiest realistic hour)

- ~0.04–0.08 rps sustained
- This validates the system under realistic beta peak load

3. Controlled spike (2–3× peak)

- ~0.10–0.25 rps sustained

This simulates onboarding bursts or short-term demand increases.
It helps us see how the system behaves beyond expected peak but still within plausible growth.

4. Stress / resilience testing

- 6–12 rps (short bursts only)

This is not expected beta traffic.

This is used to:

- Understand scaling limits
- Observe degradation behaviour
- Identify failure modes
- See whether throttling or latency spikes occur

This informs future growth planning rather than private beta readiness.

What matters most under load

The following API-driven journeys must remain stable under expected peak:

1. Initial search
2. Applying filters
3. Opening course details
4. Pagination
5. Location/autocomplete

If any of these degrade significantly, user experience and trust will be affected.

How users are expected to behave

Based on proxy data from FAC and FAA:

- Users perform between 2–6 searches per session.
- Filter changes are common.
- Pagination is common.
- Location autocomplete is likely to be heavily used.
- Searches are typically filter-heavy.

Our load model will reflect a realistic mix of:

- Initial searches
- Filtered searches
- Pagination
- Details calls
- Location lookups

This ensures we're not testing synthetic or unrealistic traffic patterns.

Performance expectations (proposed targets)

For private beta, we propose the following targets:

- Initial search: p95 ≤ 2 seconds
- Filtered search: p95 ≤ 2.5 seconds
- Pagination: p95 ≤ 1.5 seconds
- Details: p95 ≤ 1.5 seconds
- Autocomplete: p95 ≤ 1 second

Error rate under expected peak should remain below 1%.

Under stress conditions, degradation should be predictable and controlled — increased latency or throttling is acceptable; instability or cascading failure is not.

These targets are open for technical sign-off.

Spikes and growth

Traffic may increase during:

- Results day
- Campaign or comms-driven activity
- Feature launches

Stress testing helps us prepare for that future state.

Assumptions

Where real data is not yet available, volumetric assumptions are documented transparently and used for modelling only.

All assumptions are subject to Product and Technical review.

As real usage data becomes available during private beta, these figures will be revisited and refined.