To help me plan meaningful performance testing for the FEAT Search API ahead of private beta (and avoid making assumptions later), it would be really helpful to get your input on the areas below. High-level estimates are absolutely fine where exact data isn't available.

For context, performance testing will be carried out on the test environment. The search setup is intended to be broadly representative of production (similar order of magnitude of index size, same schema, embeddings, filters, AI Search configuration and caching behaviour). Any known differences or limitations will be clearly documented.

1. Purpose

- What is the main goal of this round of performance testing?

The main goal of this round of performance testing is to validate that the FEAT Search API can operate reliably and consistently under expected private beta load. This includes establishing baseline performance metrics, identifying any performance bottlenecks or degradation points, and confirming that the system behaves predictably under both normal and peak usage scenarios.

- To help me understand how to use the results for private beta, what decisions would you like performance testing to inform? For example:
    - whether we're safe to onboard an initial group of users

Performance testing will indicate whether the platform is sufficiently stable to onboard an initial group of users while maintaining a reliable and responsive search experience.

    - whether we should cap or phase traffic early on

The results will inform whether a phased onboarding approach or traffic caps are required during early private beta to manage risk and prevent performance degradation as user numbers increase.

    - whether there are any performance risks we're happy to accept at this stage

Performance testing will inform whether any performance risks can be accepted at this stage, and whether those risks are sufficiently understood and managed.

    - whether we need extra monitoring or alerts in place before opening access

Performance testing will be used to determine whether additional monitoring, alerting, or operational safeguards are required before opening private beta access, ensuring that any performance degradation or instability can be detected and addressed quickly.

2. Expected traffic levels

- What is the expected average traffic to Search?
- What is the expected peak traffic (users/min or requests/sec)?
- Do we know any upper bounds we should avoid exceeding?

The expected traffic levels outlined below are assumptions for performance testing purposes only and do not represent confirmed or actual usage figures. These assumptions are based on the understanding that the private beta access link has been shared with approximately 250 users and are intended to support realistic load modelling in the absence of validated traffic data.

Under this assumption, not all invited users are expected to be active concurrently, and usage is expected to ramp gradually rather than all users accessing the service at the same time. For test planning, it is assumed that 30–50% of invited users may be active on a typical day (75–125 users), with each active user performing 5–6 searches per session. This results in an estimated average traffic level of approximately 375–750 searches per day, or around 30–60 searches per hour during quieter periods, assuming activity is concentrated within typical waking and working hours.

Peak traffic is assumed to occur during working and school hours. For performance testing purposes, it is assumed that 10–20% of invited users may be active during the busiest hour. This would equate to approximately 25–50 users in the busiest hour, generating around 125–300 searches per hour, or approximately 0.03 – 0.08 requests per second.

To account for short-term spikes, onboarding bursts, or unexpected increases in usage, performance testing will also include an assumed upper bound of 2–3× the expected peak traffic. This results in an upper-bound test range of approximately 6–12 requests per second, which will be used to assess system behaviour under stress and inform decisions on traffic capping or phased access during private beta.

3. Usage context & existing data

- As FEAT consolidates FAC, FAA and NCS, do we have any usage stats we can use as a proxy?
    - Monthly users-
    - Searches per user (if known)
- Do we have estimated monthly users by group (e.g. young people, parents, advisers, adults returning to work)?

Assumption for FAC:

Monthly Users-

Searches per user- 4.98

Estimated users by group- FAC do not track users by group. So, our assumption is that it is most likely that the main users are young people, advisers and parents.

FAA:

Monthly users – 200,000

Searches per user – 2.01

Estimated users by group- FAA do not track users by group. So, our assumption is that it is most likely that the main users are young people, advisers and parents.

4. Search behaviour

- On average, how many searches does a user perform per session?
- How common is it for users to:
    - change filters
    - paginate results
    - re-search with new keywords?
- Do we expect autocomplete to be heavily used?
- Are searches typically keyword-only, location + distance, or filter-heavy?

Assumption for FAC

FAA:

2.01 searched per session

It is very common for users to change the filters and common for the users to move onto the next page. It is not as common for a user to research with new words as there is a job category filter.

Autocomplete is not applied in the keyword section but there is autocomplete in the where section which covers location – this would be heavily used.

Searches are filter heavy whilst using the sort option as well.

5. Critical journeys

Which of the following are most business-critical (i.e. must not degrade under load)?

- Search results load
- Applying filters
- Pagination
- Opening details
- Location / autocomplete

Search results load, Applying filters, Opening details, Pagination, Autocomplete/Location

6. Performance expectations

- What response times are acceptable for first search, filtered search, and pagination?
- What is an acceptable error rate under peak load?
- Are there any expectations around timeouts (e.g. p95 under X seconds)?

7. Session length & timing

- What is the typical session length for a FEAT search user?
    - quick lookup (30–60 seconds)
    - more exploratory (2–5 minutes)
- Do we expect usage to vary by time of day (working hours vs evenings/weekends)?

  Yes, it would be heavier usage during working and school hours and lighter during the evenings/ weekends

- Do we have a rough sense of the percentage of traffic in the busiest hour?
- If not, is it OK to use DfE GOV.UK traffic trends as a proxy?

  Yes it is ok to use DFE GOV.UK as a proxy

8. Spikes & future growth

- Are there expected campaign or comms-driven spikes (e.g. launches or announcements)?

  Yes, most likely around launches, comms for certain roles and results day

- If so, would assuming a 2×–3× multiplier on normal peak traffic be reasonable for spike testing?
- Is FEAT expected to become the primary discovery route over time?

  Yes, FEAT will be expected to become the primary discovery route
- Will traffic be replaced immediately or grow gradually?

  It will be a gradual ramp not an immediate full replacement

9. Assumptions

- Where exact data isn't available, is it acceptable for QA to document reasonable assumptions for volumetrics and load modelling, subject to PM/Product sign-off?

Yes

Thanks — this will help me size the tests realistically and make sure the results are meaningful for the private beta.