Streaming

- Disabling streaming
- When to use streaming vs non-streaming

Certain API endpoints stream responses by default, such as `/api/generate`. These responses are provided in the newline-delimited JSON format (i.e. the `application/x-ndjson` content type). For example:

```
{"model":"gemma4","created_at":"2025-10-26T17:15:24.097767Z","response":"That","done":false}
{"model":"gemma4","created_at":"2025-10-26T17:15:24.109172Z","response":"'","done":false}
{"model":"gemma4","created_at":"2025-10-26T17:15:24.121485Z","response":"s","done":false}
{"model":"gemma4","created_at":"2025-10-26T17:15:24.132802Z","response":" a","done":false}
{"model":"gemma4","created_at":"2025-10-26T17:15:24.143931Z","response":"
fantastic","done":false}
{"model":"gemma4","created_at":"2025-10-26T17:15:24.155176Z","response":"
question","done":false}
{"model":"gemma4","created_at":"2025-10-26T17:15:24.166576Z","response":"!","done":true,
"done_reason": "stop"}
```

Disabling streaming

Streaming can be disabled by providing `{"stream": false}` in the request body for any endpoint that support streaming. This will cause responses to be returned in the `application/json` format instead:

```
{"model":"gemma4","created_at":"2025-10-26T17:15:24.166576Z","response":"That's a fantastic
question!","done":true}
```

When to use streaming vs non-streaming

Streaming (default):

- Real-time response generation
- Lower perceived latency
- Better for long generations

Non-streaming:

- Simpler to process
- Better for short responses, or structured outputs
- Easier to handle in some applications

□□□ #2
□ □□□ □□□ 2026-10-31 13:16:31 UTC
□ □□□ □□□ 2026-10-31 13:16:32 UTC