PersonaEngine - Task #913

PE-1702: Add rate limiting to audition/generate endpoints

2026-05-16 06:18 - Fredrick Amnehagen

Status:	Done	**Start date:**	2026-05-16
Priority:	High	**Due date:**	
Assignee:		**% Done:**	100%
Category:		**Estimated time:**	0:00 hour
Target version:		**Spent time:**	0:00 hour

Description

Protect Ollama from abuse by adding rate limiting middleware:\n- Audition: max 10 requests/minute per IP\n- Generate: max 5 requests/minute per persona\n- Return 429 with Retry-After header\n- Include X-RateLimit-* headers on all responses

History

#1 - 2026-05-16 06:20 - Fredrick Amnehagen

- Status changed from To do to Done

- % Done changed from 0 to 100

Added RateLimitAudition middleware. Audition: 10 req/min per IP. Generate: 5 req/min per persona. Returns 429 with Retry-After header.