MedScraper AI – Back-Dated Commit Timeline

Project window: 2 September 2024 ➔ 7 February 2025 (≈ 22 weeks, ~150 commits)

> *Tip → Push to* `main` *(or* `develop`*) on the stated dates. Use* `--date` *flag to back-date if rewriting history, e.g.* `GIT_COMMITTER_DATE="2024-09-02 09:30" git commit --amend --no-edit --date "2024-09-02 09:30"`

September 2024 (Foundation – 43 commits)

Date	Commit message
2024-09-02	Initial commit — create FastAPI project skeleton & `pyproject.toml`
2024-09-02	Add Dockerfile + docker-compose with PostgreSQL & Redis
2024-09-02	Configure pre-commit, black, flake8, mypy linters
2024-09-03	Add basic `Article` Pydantic model & SQLAlchemy ORM definition
2024-09-03	Set up Alembic migration environment & initial revision
2024-09-03	Add health-check endpoint `/health`
2024-09-04	Implement PostgreSQL connection pool via asyncpg
2024-09-04	Wire first Alembic autogenerate migration to create `articles` table
2024-09-04	Add CI workflow: lint + unit-tests on GitHub Actions
2024-09-05	Create Redis & Celery boilerplate with simple “ping” task
2024-09-05	Add Flower for Celery monitoring (docker-compose profile)
2024-09-05	Write README project overview & local setup instructions
2024-09-06	Add utility logging wrapper, structured JSON logs
2024-09-06	Add SPA placeholder for Swagger docs customization
2024-09-06	Set up `.env.example` and secrets handling via Pydantic Settings
2024-09-07	Implement async scraper service layout (interface only)
2024-09-07	Add httpx async client with retry + timeout middleware
2024-09-08	Create parser helpers: BeautifulSoup + lxml abstractions
2024-09-08	Add RSS feedparser utility function
2024-09-08	Unit-test RSS utility with NIH feed sample

Date	Commit message
2024-09-09	Implement Medical News Today scraper (HTML)
2024-09-09	Persist scraped items to DB via repository layer
2024-09-09	Add duplicate detection by URL hash
2024-09-10	Write Celery task `scrape_source` & schedule via celery-beat
2024-09-10	Add initial basic keyword filter for medical content
2024-09-10	Integrate slowapi rate-limiting on scraper requests
2024-09-11	Add WebMD RSS scraper
2024-09-11	Refactor scraper base class for pluggable sources
2024-09-11	Unit tests for duplicate detection logic
2024-09-12	Add Reuters Health RSS scraper
2024-09-12	Add tag/section filter for non-medical pruning
2024-09-13	Refactor settings into `settings.py` with typed env config
2024-09-13	Add GitHub Action: build & push Docker image on main
2024-09-13	Write docs: scraping architecture diagrams (Markdown)
2024-09-14	Add `/articles` GET endpoint with pagination
2024-09-14	Add SQL index on `published_at` column
2024-09-15	Add pytest-asyncio tests for `/articles` route
2024-09-15	Configure Alembic autogenerate in CI
2024-09-16	Add `/refresh` POST endpoint to queue scrape jobs
2024-09-16	Implement JWT auth stub (no roles yet)
2024-09-17	Add error handling middleware & global exception model
2024-09-17	Improve logging format for Celery tasks
2024-09-18	Docs: update README with API usage examples
2024-09-18	Tag v0.1.0 – Core data ingestion & API draft

October 2024 (NLP & Business Logic – 38 commits)

Date	Commit message
2024-10-01	Add HuggingFace transformers & torch to requirements

Date	Commit message
2024-10-01	Implement summarization service with BART-large-cnn
2024-10-02	Add Celery task `generate_summary` per article
2024-10-02	Store summary text in DB migration
2024-10-02	Unit test summarization length constraints
2024-10-03	Integrate spaCy + PubMedBERT NER pipeline
2024-10-03	Persist entities as JSONB column + migration
2024-10-03	Add test fixture for NER parsing
2024-10-04	Implement sentiment analysis with twitter-roberta
2024-10-04	Cache HF models on startup – lazy singleton
2024-10-05	Refactor NLP tasks into `nlp_worker.py`
2024-10-05	Add Celery chord: scrape ➜ nlp summarization ➜ flag alerts
2024-10-06	Add classification categories (breakthrough / recall / advisory)
2024-10-06	Create simple rules engine; add mapping table
2024-10-07	Add credibility score field; naive source weighting
2024-10-07	Write unit tests for credibility scoring
2024-10-08	Expose `/summary/{id}` endpoint
2024-10-08	Add NER/summary/sentiment to `/articles` response model
2024-10-09	Update OpenAPI docs with examples
2024-10-09	Improve duplicate detection with Levenshtein title distance
2024-10-10	Add periodic cleanup task: delete stale temp files
2024-10-10	Add rate-limit headers on public endpoints
2024-10-11	Integrate Sentry SDK for error tracking
2024-10-11	Add pytest-cov & enforce 80% coverage gate
2024-10-12	Add black + isort auto-format CI step
2024-10-12	Refactor settings to nested BaseSettings classes
2024-10-13	Docs: NLP pipeline architecture diagram
2024-10-13	Tag v0.2.0 – NLP MVP complete
2024-10-15	Add Prometheus exporter optional endpoint

Date	Commit message
2024-10-15	Add pagination metadata headers
2024-10-16	Optimize db queries – eager loading for joins
2024-10-16	Add index on `credibility_score`
2024-10-17	Implement bloom filter for URL duplicates – memory cache
2024-10-18	Stress test scraper with 10k articles fixture
2024-10-18	Fix: RSS parser timezone normalization bug
2024-10-19	Write ADR document for NLP model choices
2024-10-20	Add `/metrics` Prometheus endpoint (optional feature flag)
2024-10-21	Docs: update README badges & coverage shield
2024-10-21	Add Makefile for common dev tasks
2024-10-22	Tag v0.3.0 – Performance & ops improvements

November 2024 (Alerts & Deployment – 30 commits)

Date	Commit message
2024-11-01	Add alert rules engine infrastructure
2024-11-01	Implement keyword alert: “outbreak” & “recall”
2024-11-02	Add webhook dispatcher (generic POST) service
2024-11-02	Write unit tests for webhook payload validation
2024-11-03	Add Slack webhook example integration
2024-11-03	Add environment-driven alert keyword list
2024-11-04	Implement user preferences table & CRUD
2024-11-04	Add `/alerts/preferences` endpoints (auth required)
2024-11-05	Integrate Celery beat schedule: run alert scan hourly
2024-11-05	Add Flower auth & reverse proxy config
2024-11-06	Docker: switch to python:3.11-slim base image
2024-11-06	Multi-stage build to shrink final image
2024-11-07	Add Nginx reverse proxy config + Docker compose prod
2024-11-07	Add Gunicorn/uvicorn worker setup

Date	Commit message
2024-11-08	Terraform scaffold for AWS networking & RDS
2024-11-08	Add GitHub Actions deploy to ECR + ECS task def
2024-11-09	Docs: deployment guide for AWS Fargate
2024-11-10	Add CloudWatch log driver config in task def
2024-11-10	Set up AWS Secrets Manager integration
2024-11-11	Add ALB target group & HTTPS cert via ACM
2024-11-12	Smoke test ECS deployment – first green run
2024-11-12	Fix: celery worker memory limits in task definition
2024-11-13	Add auto-scaling policy on CPU > 60%
2024-11-14	Integrate uptime monitor ping (UptimeRobot)
2024-11-14	Add /health → ALB health-check path
2024-11-15	Docs: architecture diagram – AWS deployment
2024-11-15	Tag v0.4.0 – Alerts + initial production deploy
2024-11-18	Add S3 log archive lifecycle policy TF
2024-11-18	Set up E2E test pipeline staging ➜ prod
2024-11-19	Optimize summarization: batch process mode
2024-11-19	Reduce container size by pruning build deps
2024-11-20	Docs: Post-mortem template added

December 2024 (Monitoring, Security, Polish – 23 commits)

Date	Commit message
2024-12-02	Add CloudWatch custom metric: scrape latency
2024-12-02	Grafana dashboard JSON export
2024-12-03	Add Sentry performance tracing middleware
2024-12-03	Enable CORS middleware with env whitelist
2024-12-04	Add rate limit redis backend for public API
2024-12-04	Add pytest-postgresql for DB integration tests
2024-12-05	Add per-source success/error metrics

Date	Commit message
2024-12-05	Tune Celery concurrency vs threads in prod
2024-12-06	Upgrade dependencies; fix newly surfaced mypy errors
2024-12-06	Add dependabot config file
2024-12-07	Doc: contributing guide + code of conduct
2024-12-07	Refactor alert keyword matching → regex per locale
2024-12-08	Add locale field to Article model + migration
2024-12-08	Implement language detection fallback
2024-12-09	Set up feature flags via environment variables
2024-12-10	Add white-label mode for partner deployments
2024-12-11	Harden security headers in Nginx conf
2024-12-11	Tag v1.0.0 – MVP feature-complete
2024-12-15	Benchmark: 50 req/s sustained load test results
2024-12-15	Docs: performance benchmark results page
2024-12-18	Refactor README → docs site (MkDocs)
2024-12-18	Add GitHub pages action for docs deploy
2024-12-20	Holiday patch: suppress alerts on low severity during holidays

January 2025 (Refinement & Stretch Goals – 16 commits)

Date	Commit message
2025-01-06	Add MLflow tracking for experiment logging
2025-01-06	Integrate caching layer for summarization results
2025-01-07	Add advanced duplicate detection via SimHash
2025-01-07	Implement full-text search with pg_trgm extension
2025-01-08	Add `/search` endpoint with ranking by recency + score
2025-01-08	Docs: search endpoint usage examples
2025-01-09	Add dark mode theme to docs site
2025-01-10	Create Helm chart for k8s alternative deployment
2025-01-10	Add GitHub Action publish Helm chart package

Date	Commit message
2025-01-13	Refactor settings to dynaconf (experimental)
2025-01-14	Add server-side metrics export in OpenTelemetry
2025-01-15	Stress test: k6 cloud run 100 VU
2025-01-16	Fix: OpenTelemetry context propagation bug
2025-01-17	Docs: 2025 roadmap draft
2025-01-20	Tag v1.1.0 – Search & observability enhancements
2025-01-31	Chore: year-end dependency updates + LICENSE review

February 2025 (Wrap-Up – 4 commits)

Date	Commit message
2025-02-03	Add retrospective ADR for MVP lessons learned
2025-02-04	Remove deprecated NLP models; pin HF model versions
2025-02-05	Docs: publish v1.1.0 release notes
2025-02-07	Tag v1.1.1 – Final polish before hand-off