SYNAPSE TO PORTFOLIO

Technical Documentation

Author: Ivan Ivanov
Date: 2026-04-15

This document describes the architecture, implementation, and operational procedures for the Synapse to Portfolio system. The system automates the transformation of raw research notes into structured, publication-ready portfolio entries. It integrates a knowledge graph, a natural language processing pipeline, and a static site generator into a single coherent workflow.

The document is intended for engineers who will maintain or extend the system, as well as for researchers who want to understand how their notes are processed. It assumes familiarity with Python, basic graph theory, and command-line tools.

Introduction

Researchers accumulate notes in many formats: Markdown files, PDF annotations, spreadsheet rows, and bibliography entries. Turning these scattered artifacts into a coherent portfolio is time-consuming and error-prone. Synapse to Portfolio addresses this problem by ingesting all of these sources, linking them through a shared knowledge graph, and producing a polished website with minimal manual intervention.

The system was motivated by the author's own experience preparing annual reports. Each year, hours were spent copying results from notebooks into slides and then into a personal website. Automating this pipeline freed that time for actual research.

Goals and Non-Goals

The primary goal is to reduce the effort required to maintain a research portfolio. A secondary goal is to preserve provenance: every claim on the website must be traceable back to the original note that produced it.

The system is not intended to replace human judgment. It will not decide what is worth publishing, and it will not write prose on the author's behalf. It only organizes and presents what the author has already written.

Architecture Overview

The system is organized as a pipeline with four stages: ingestion, enrichment, graph construction, and rendering. Each stage reads from the previous stage's output and writes to a well-defined intermediate format. This makes it possible to inspect or replace any stage independently.

Ingestion watches a directory for new or modified files and converts them into a common internal representation. Enrichment runs named-entity recognition and topic modeling over the text. Graph construction links entities and documents into a knowledge graph stored in SQLite. Rendering queries the graph and produces static HTML.

Component Diagram

The following diagram shows the main components and the data flows between them. Boxes represent components; arrows represent data flows. Dashed arrows indicate optional flows that are only used when a particular feature is enabled.

Ingestion and Parsing

The ingestion stage is implemented in the `synapse.ingest` module. It uses the `watchdog` library to receive filesystem events and a registry of parsers to handle different file types. Each parser is a function that takes a path and returns a list of documents.

Markdown files are parsed with `markdown-it-py` and front matter is extracted with `python-frontmatter`. PDF files are converted to text with `pdfminer.six`. Bibliography entries are read from BibTeX files with `bibtexparser`. Spreadsheet rows are read with `openpyxl`.

Supported Formats

Format	Extension	Parser	Notes
Markdown	`.md`	`parse_markdown`	Front matter supported
PDF	`.pdf`	`parse_pdf`	Text only, no figures
BibTeX	`.bib`	`parse_bibtex`	One entry per document
Spreadsheet	`.xlsx`	`parse_xlsx`	One row per document
Plain text	`.txt`	`parse_text`	Fallback parser

Incremental Updates

To avoid reprocessing the entire corpus on every change, the ingestion stage stores a hash of each file in a table called `ingest_state`. When a file event arrives, the new hash is compared with the stored one. If they match, the file is skipped. This makes the watcher cheap to run continuously.

Enrichment

Enrichment adds structure to the raw text. It runs named-entity recognition with `spaCy`, topic modeling with `gensim`, and keyword extraction with `YAKE`. The results are stored as annotations attached to each document.

The enrichment stage is the most computationally expensive part of the pipeline. It can be disabled with the `--no-enrich` flag when only a quick rebuild is needed.

Named Entities

Named entities are grouped into the following categories: people, organizations, locations, dates, and domain-specific terms. The domain-specific category is populated from a user-maintained list in `config/terms.txt`.

Topic Modeling

Topic modeling uses latent Dirichlet allocation with a default of 20 topics. The number of topics can be changed in the configuration file. Each document receives a distribution over topics, and the dominant topic is stored as a single label for convenience.

Knowledge Graph

The knowledge graph is stored in SQLite using two tables: `nodes` and `edges`. Nodes represent documents, entities, and topics. Edges represent relationships such as "mentions", "belongs to", and "cites". The schema is deliberately simple so that it can be queried with plain SQL.

Schema

```
CREATE TABLE nodes (
    id INTEGER PRIMARY KEY,
    kind TEXT NOT NULL,
    label TEXT NOT NULL,
    payload TEXT
);

CREATE TABLE edges (
    src INTEGER NOT NULL,
    dst INTEGER NOT NULL,
    kind TEXT NOT NULL,
    weight REAL DEFAULT 1.0
);
```

Queries

Common queries are wrapped in helper functions in `synapse.graph.queries`. For example, `neighbors(node_id, kind)` returns all nodes connected to a given node by edges of a given kind. More complex queries can be written directly against the database.

Rendering

Rendering uses `Jinja2` templates and `Tailwind CSS`. The renderer walks the graph, groups documents by topic, and writes one HTML page per topic plus an index page. Static assets are copied into the output directory unchanged.

The templates live in `templates/` and can be overridden by placing files with the same name in a user directory. This allows customization without forking the repository.

Theming

Themes are defined as CSS variables in `templates/theme.css`. Switching themes is a matter of changing the `data-theme` attribute on the root element. Two themes are shipped by default: `light` and `dark`.

Configuration

Configuration is read from `synapse.toml` in the working directory. The file is optional; sensible defaults are used when it is absent. Environment variables with the prefix `SYNAPSE_` override values from the file.

Key	Default	Description
`watch.dir`	`./notes`	Directory to watch for changes
`enrich.topics`	`20`	Number of LDA topics
`render.out`	`./site`	Output directory for HTML
`render.theme`	`light`	Theme name
`graph.db`	`./synapse.db`	SQLite database path

Operations

The system is operated through a single command-line tool called `synapse`. Subcommands cover ingestion, enrichment, rendering, and maintenance tasks. Run `synapse --help` for the full list.

Backups

The SQLite database is the only stateful component. Backups can be taken with `synapse backup`, which writes a timestamped copy to `./backups/`. Restoring is a matter of replacing the database file.

Monitoring

The watcher writes a heartbeat file every minute. A simple cron job can check that the file is recent and alert if it is not. More elaborate monitoring is left to the operator.

Limitations

The system has several known limitations. Figures in PDFs are ignored. Tables are flattened to plain text. Non-English documents are processed but the topic model is tuned for English. These limitations are acceptable for the author's use case but may not be for others.

Future Work

Future work includes figure extraction, multilingual topic models, and a web interface for browsing the graph. Contributions are welcome; see `CONTRIBUTING.md` for guidelines.

Conclusion

Synapse to Portfolio turns a scattered collection of notes into a structured, searchable, and publishable portfolio with minimal effort. The pipeline is simple enough to understand in an afternoon and flexible enough to adapt to many workflows.

The author hopes that the system will save others the same hours it saved him, and that the open-source release will attract contributions that make it better than he could alone.

Appendix A: Full Schema

The full schema, including indexes and foreign keys, is reproduced here for reference. It is also available in `synapse/graph/schema.sql`.

```
CREATE TABLE nodes (
    id INTEGER PRIMARY KEY,
    kind TEXT NOT NULL,
    label TEXT NOT NULL,
    payload TEXT
);
CREATE INDEX nodes_kind ON nodes(kind);
CREATE INDEX nodes_label ON nodes(label);

CREATE TABLE edges (
    src INTEGER NOT NULL REFERENCES nodes(id),
    dst INTEGER NOT NULL REFERENCES nodes(id),
    kind TEXT NOT NULL,
    weight REAL DEFAULT 1.0
);
CREATE INDEX edges_src ON edges(src);
CREATE INDEX edges_dst ON edges(dst);
```

Appendix B: Glossary

Ingestion — the process of reading files and converting them into documents.

Enrichment — the process of adding structure to documents through NLP.

Graph — the knowledge graph stored in SQLite.

Rendering — the process of producing HTML from the graph.

Watcher — the component that monitors the filesystem for changes.

Front matter — YAML metadata at the top of a Markdown file.

Topic — a cluster of words produced by the topic model.

Entity — a named person, organization, location, date, or term.

Node — a row in the `nodes` table.

Edge — a row in the `edges` table.

Heartbeat — a file touched periodically to indicate liveness.

Backup — a timestamped copy of the database.

Theme — a set of CSS variables controlling the appearance.

Template — a Jinja2 file used to render HTML.

Static site — the output directory served by any web server.

Provenance — the chain from a published claim back to its source note.

Operator — the person who runs the system.

Author — the person who writes the notes.

Reader — the person who visits the website.

Maintainer — the person who changes the code.

Contributor — anyone who submits a patch.

Release — a tagged version of the code.

Changelog — the file listing changes between releases.

License — the terms under which the code is distributed.

Issue — a bug report or feature request.

Pull request — a proposed change to the code.

Continuous integration — automated tests run on every change.

Linter — a tool that checks code style.

Formatter — a tool that rewrites code to a canonical style.

Type checker — a tool that verifies type annotations.

Test suite — the collection of automated tests.

Coverage — the fraction of code exercised by tests.

Benchmark — a measurement of performance.

Profile — a detailed breakdown of where time is spent.

Cache — a store of previously computed results.

Index — a data structure that speeds up queries.

Migration — a change to the database schema.

Rollback — undoing a migration.

Dependency — a third-party library used by the code.

Lockfile — a file pinning exact dependency versions.

Virtual environment — an isolated Python installation.

Container — a packaged runtime environment.

Image — a snapshot used to create containers.

Volume — persistent storage for containers.

Port — a network endpoint exposed by a container.

Environment variable — a configuration value passed to a process.

Secret — a sensitive value such as a password or token.

Log — a record of events written by the system.

Log level — the severity of a log message.

Structured logging — logs written as machine-readable records.

Metric — a numeric measurement tracked over time.

Dashboard — a visual display of metrics.

Alert — a notification triggered by a metric.

Incident — an unplanned interruption of service.

Postmortem — a written analysis of an incident.

Runbook — a step-by-step guide for operators.

On-call — the person responsible for responding to alerts.

Service level objective — a target for reliability.

Error budget — the allowed amount of unreliability.

Rate limit — a cap on how often something may happen.

Retry — attempting an operation again after failure.

Timeout — a limit on how long to wait.

Circuit breaker — a mechanism that stops calling a failing service.

Idempotent — safe to perform more than once.

Transaction — a group of operations that succeed or fail together.

Deadlock — two processes waiting for each other forever.

Race condition — a bug caused by unexpected timing.

Thread — a unit of concurrent execution.

Process — a running program with its own memory.

Coroutine — a function that can pause and resume.

Event loop — the scheduler for asynchronous code.

Future — a placeholder for a result not yet available.

Callback — a function passed to be called later.

Generator — a function that yields values one at a time.

Iterator — an object that produces values on demand.

Context manager — an object used with `with` to manage resources.

Decorator — a function that wraps another function.

Metaclass — a class that creates classes.

Descriptor — an object that customizes attribute access.

Slot — a fixed attribute declaration to save memory.

Weak reference — a reference that does not prevent collection.

Garbage collection — automatic memory reclamation.

Memory leak — memory that is never reclaimed.

Profiler — a tool that measures where time is spent.

Flame graph — a visualization of profiler output.

Hot path — code that runs frequently and matters for speed.

Premature optimization — optimizing before measuring.

Refactoring — changing code without changing behavior.

Technical debt — shortcuts that will cost time later.

Code review — another person reading your change.

Pair programming — two people working on one change.

Mob programming — a group working on one change.

Trunk-based development — small changes merged often.

Feature branch — a long-lived branch for one feature.

Merge conflict — two changes that cannot be combined automatically.

Rebase — replaying commits on top of another branch.

Squash — combining commits into one.

Cherry-pick — applying a single commit elsewhere.

Tag — a named pointer to a commit.

Semantic versioning — major.minor.patch version numbers.

Deprecation — marking something for future removal.

Backward compatibility — old code still works.

Forward compatibility — new data still works with old code.

Schema evolution — changing the database structure over time.

Data migration — moving data to match a new schema.

Seed data — initial data for development and tests.

Fixture — test data prepared in advance.

Mock — a fake object used in tests.

Stub — a minimal implementation used in tests.

Integration test — a test that exercises several components.

Unit test — a test that exercises one function.

End-to-end test — a test that exercises the whole system.

Smoke test — a quick test that the system starts.

Regression test — a test added after a bug is fixed.

Property-based test — a test that checks invariants on random inputs.

Fuzzing — feeding random inputs to find crashes.

Static analysis — checking code without running it.

Dynamic analysis — checking code while it runs.

Sandbox — an isolated environment for untrusted code.

Permission — a right to perform an action.

Role — a named set of permissions.

Authentication — proving who you are.

Authorization — deciding what you may do.

Session — a period of interaction with a user.

Cookie — a small piece of data stored by the browser.

Token — a string proving identity or permission.

Hash — a fixed-size fingerprint of data.

Salt — random data mixed into a password hash.

Encryption — making data unreadable without a key.

Key — a secret used for encryption or signing.

Certificate — a public key with an identity.

Handshake — the setup of a secure connection.

Protocol — a set of rules for communication.

Packet — a unit of network data.

Latency — the delay before a response.

Throughput — the amount of work per unit time.

Bandwidth — the capacity of a network link.

Load balancer — a component that distributes requests.

Reverse proxy — a server in front of application servers.

Cache invalidation — removing stale cached data.

Eventual consistency — replicas converge over time.

Strong consistency — every read sees the latest write.

Quorum — a majority of replicas.

Leader election — choosing one node to coordinate.

Heartbeat — a periodic signal that a node is alive.

Split brain — two nodes both think they are leader.

Backpressure — slowing producers when consumers are busy.

Queue — a buffer between producers and consumers.

Dead letter queue — where failed messages go.

Publish-subscribe — one message to many listeners.

Request-response — one message and one reply.

Streaming — data sent as a continuous flow.

Batch — data processed in large chunks.

Checkpoint — a saved position in a stream.

Offset — a position in a stream.

Partition — a shard of a stream.

Consumer group — workers sharing a stream.

Exactly once — each message processed one time.

At least once — messages may be duplicated.

At most once — messages may be lost.

Schema registry — a service for message schemas.

Serialization — turning data into bytes.

Deserialization — turning bytes into data.

Binary format — a compact machine-readable encoding.

Text format — a human-readable encoding.

Compression — reducing the size of data.

Checksum — a value to detect corruption.

Versioning — tracking changes over time.

Rollout — deploying a change gradually.

Canary — a small first deployment.

Blue-green — two identical environments for switching.

Feature flag — a switch to enable code at runtime.

Kill switch — a flag to disable a feature quickly.

Dark launch — running code without showing results.

A/B test — comparing two variants.

Cohort — a group of users.

Funnel — a sequence of steps users take.

Conversion — reaching a desired step.

Retention — users returning over time.

Churn — users leaving over time.

Engagement — how much users interact.

Onboarding — the first experience of a new user.

Documentation — written explanations for users.

Tutorial — a guided lesson.

Reference — a complete description of an interface.

Example — a small working program.

Cookbook — a collection of recipes for tasks.

Changelog — a record of changes between releases.

Release notes — a human-written summary of a release.

Migration guide — how to move between versions.

Deprecation notice — a warning that something will be removed.

Support — help for users with problems.

Community — the people who use and discuss the project.

Governance — how decisions are made.

Roadmap — a plan for future work.

Vision — the long-term goal of the project.

Mission — why the project exists.

Values — principles guiding the project.

Code of conduct — rules for behavior in the community.

Maintainer — someone responsible for the project.

Contributor — someone who has submitted a change.

Author — someone who has written a change.

Reviewer — someone who has reviewed a change.

Committer — someone who can merge changes.

Owner — someone who can grant permissions.

Bus factor — how many people must leave before the project stalls.

Onboarding — helping a new contributor get started.

Mentorship — guiding a contributor over time.

Recognition — thanking contributors publicly.

Sponsorship — funding for the project.

Donation — money given to the project.

Grant — money awarded for specific work.

Bounty — money offered for a specific fix.

Issue triage — sorting and labeling new issues.

Label — a tag on an issue or pull request.

Milestone — a group of issues for a release.

Project board — a visual view of work in progress.

Standup — a short daily meeting.

Retrospective — a meeting to improve the process.

Planning — deciding what to work on next.

Estimation — guessing how long work will take.

Velocity — how much work is done per iteration.

Burndown — remaining work over time.

Definition of done — when a task is finished.

Acceptance criteria — what a feature must do.

User story — a description of a feature from a user's view.

Epic — a large user story.

Task — a small piece of work.

Subtask — a piece of a task.

Dependency — work that must finish before other work.

Blocker — something preventing progress.

Workaround — a temporary fix for a problem.

Root cause — the underlying reason for a problem.

Symptom — the visible effect of a problem.

Reproduction — steps to make a problem happen.

Minimal example — the smallest code that shows a problem.

Bisect — finding the commit that introduced a problem.

Revert — undoing a commit.

Hotfix — an urgent fix released quickly.

Patch — a small change to fix a problem.

Backport — applying a fix to an older release.

Forward port — applying a fix to a newer release.

Vendor — a third party providing software.

Open source — software with public source code.

Proprietary — software with restricted source code.

Free software — software with the four freedoms.

Public domain — work with no copyright.

Attribution — crediting the author.

Boilerplate — repetitive standard text.

Template — a reusable pattern for new files.

Scaffold — generated starting code.

Generator — a tool that produces code.

Code generation — writing code with code.

Macro — a compile-time code expansion.

Preprocessor — a tool that runs before the compiler.

Compiler — a tool that translates code to machine code.

Interpreter — a tool that runs code directly.

Bytecode — an intermediate machine-like format.

Virtual machine — software that runs bytecode.

Just-in-time compilation — compiling while running.

Ahead-of-time compilation — compiling before running.

Linker — a tool that combines compiled pieces.

Loader — a tool that puts a program in memory.

Executable — a file that can be run directly.

Shebang — the `#!` line telling the system how to run a script.

Environment — the variables and settings a process sees.

Working directory — the default location for relative paths.

Path — a location in the filesystem.

Absolute path — a path from the root.

Relative path — a path from the working directory.

Symlink — a file pointing to another file.

Hardlink — two names for the same file.

Permission bits — read, write, and execute flags.

Owner — the user who controls a file.

Group — a set of users sharing access.

Umask — the default permissions for new files.

Signal — a message sent to a process.

Interrupt — a signal asking a process to stop.

Kill — a signal forcing a process to stop.

Daemon — a background service process.

Supervisor — a process that restarts services.

Init system — the first process started by the kernel.

Cron — a scheduler for recurring jobs.

Pipeline — a sequence of processing stages.

Stage — one step in a pipeline.

Artifact — a file produced by a build.

Build — the process of producing artifacts.

Clean build — a build from scratch.

Incremental build — a build reusing previous results.

Dependency graph — what depends on what.

Topological order — an order respecting dependencies.

Cycle — a dependency that loops back.

Vendoring — copying dependencies into the repository.

Lockfile — a file pinning exact dependency versions.

Reproducible build — a build that yields identical output.

Hermetic build — a build isolated from the environment.

Cache — a store of previous build results.

Remote cache — a cache shared between machines.

Distributed build — a build spread across machines.

Test runner — a tool that executes tests.

Test discovery — finding tests automatically.

Test fixture — setup and teardown for tests.

Parametrized test — a test run with many inputs.

Snapshot test — a test comparing against stored output.

Golden file — the stored expected output.

Flaky test — a test that sometimes fails.

Quarantine — isolating flaky tests.

Coverage report — a summary of tested code.

Branch coverage — whether each branch was taken.

Line coverage — whether each line was run.

Mutation testing — checking tests by breaking code.

Benchmark suite — a collection of performance tests.

Regression — a performance or correctness loss.

Baseline — a previous measurement to compare against.

Noise — variation in measurements.

Warm-up — running code before measuring it.

Cold start — the first run with nothing cached.

Steady state — after warm-up and caches fill.

Allocation — reserving memory.

Deallocation — releasing memory.

Memory pool — a reusable block of memory.

Arena — memory freed all at once.

Stack — memory for function calls.

Heap — memory for dynamic allocation.

Pointer — a value that refers to memory.

Reference — a safe pointer.

Ownership — who is responsible for freeing memory.

Borrow — temporary access to memory.

Lifetime — how long memory is valid.

Null — a pointer to nothing.

Dangling pointer — a pointer to freed memory.

Buffer overflow — writing past the end of memory.

Use after free — using memory that was freed.

Double free — freeing memory twice.

Memory safety — avoiding these bugs.

Sanitizer — a tool that detects memory bugs.

Valgrind — a tool for finding memory errors.

Profiler — a tool that measures where time is spent.

Sampling profiler — a profiler that interrupts periodically.

Instrumenting profiler — a profiler that records every call.

Call graph — who calls whom.

Self time — time spent in a function itself.

Total time — time including callees.

Hot function — a function that runs often.

Inlining — copying a function into its caller.

Tail call — a call as the last action.

Recursion — a function calling itself.

Iteration — looping instead of recursing.

Memoization — caching function results.

Dynamic programming — solving by remembering subproblems.

Greedy — always taking the best local choice.

Backtracking — undoing choices that fail.

Search — finding a value in data.

Sort — putting data in order.

Hash table — a fast key-value structure.

Tree — a hierarchical structure.

Balanced tree — a tree kept shallow.

Trie — a tree of prefixes.

Graph — nodes connected by edges.

Breadth-first search — exploring level by level.

Depth-first search — exploring one path deeply.

Shortest path — the cheapest route between nodes.

Minimum spanning tree — the cheapest connecting edges.

Topological sort — ordering respecting dependencies.

Strongly connected component — a mutually reachable group.

Cut vertex — a node whose removal disconnects.

Bridge — an edge whose removal disconnects.

Flow — how much can move through a network.

Matching — pairing nodes without sharing.

Coloring — labeling nodes so neighbors differ.

Planar graph — one drawable without crossings.

Tree decomposition — a tree describing a graph.

Treewidth — how tree-like a graph is.

Approximation — a near-optimal answer.

Heuristic — a rule of thumb.

Randomized algorithm — one using random choices.

Deterministic — always the same result.

Probabilistic — correct with high probability.

Hash function — a function producing fingerprints.

Collision — two inputs with the same hash.

Bloom filter — a set that may say yes falsely.

Sketch — a small summary of a large stream.

Reservoir sampling — picking a random sample from a stream.

Consistent hashing — hashing that survives resizing.

Rendezvous hashing — another resizing-friendly hash.

Merkle tree — a tree of hashes for verification.

Content addressing — naming data by its hash.

Deduplication — storing identical data once.

Compression — reducing the size of data.

Lossless — compression that loses nothing.

Lossy — compression that discards detail.

Entropy — how unpredictable data is.

Huffman coding — variable-length prefix codes.

Run-length encoding — compressing repeated values.

Dictionary coding — replacing phrases with codes.

Block — a fixed-size chunk of data.

Stream — data of unknown length.

Framing — marking where messages start and end.

Checksum — a value to detect corruption.

Error correction — repairing corrupted data.

Redundancy — extra data for safety.

Parity — a simple redundancy bit.

Erasure code — redundancy that survives lost pieces.

Replication — keeping copies on several machines.

Sharding — splitting data across machines.

Rebalancing — moving shards to even out load.

Hot spot — a shard receiving too much traffic.

Tombstone — a marker for deleted data.

Compaction — rewriting data to reclaim space.

Write-ahead log — a log written before changes.

Snapshot — a full copy of state at a point.

Recovery — restoring state after a crash.

Durability — data surviving crashes.

Fsync — forcing data to disk.

Page cache — the operating system's memory cache.

Direct I/O — bypassing the page cache.

Memory-mapped file — a file accessed like memory.

Sequential access — reading in order.

Random access — reading anywhere.

Seek — moving to a position.

Block device — hardware storing fixed blocks.

Solid state drive — flash storage.

Hard disk drive — spinning magnetic storage.

Network filesystem — files shared over a network.

Object storage — storing whole objects by key.

Key-value store — a database of pairs.

Document store — a database of documents.

Relational database — tables with joins.

Query planner — choosing how to run a query.

Index — a structure speeding up lookups.

Join — combining rows from two tables.

Transaction — a group of operations that succeed or fail together.

Isolation — transactions not seeing each other.

Serializable — as if transactions ran one at a time.

Snapshot isolation — each transaction sees a frozen view.

Optimistic concurrency — checking for conflicts at the end.

Pessimistic concurrency — locking before changing.

Lock — exclusive access to a resource.

Shared lock — many readers at once.

Exclusive lock — one writer alone.

Deadlock — two processes waiting for each other forever.

Livelock — processes busy but making no progress.

Starvation — a process never getting its turn.

Fairness — everyone getting a turn.

Priority — some work going first.

Preemption — taking the processor away.

Cooperative — processes yielding voluntarily.

Scheduler — the component deciding what runs.

Context switch — changing the running process.

Thread pool — reusable worker threads.

Work stealing — idle workers taking from busy ones.

Parallelism — doing many things at once.

Concurrency — dealing with many things at once.

Race condition — a bug caused by unexpected timing.

Memory barrier — ordering memory operations.

Atomic — indivisible operation.

Compare-and-swap — atomic conditional update.

Lock-free — progress without locks.

Wait-free — every thread finishes in bounded steps.

Transactional memory — memory with transactions.

Message passing — communicating by sending messages.

Shared memory — communicating through memory.

Actor — an isolated unit with a mailbox.

Mailbox — a queue of incoming messages.

Supervision — restarting failed actors.

Let it crash — failing fast and recovering.

Idempotent — safe to repeat.

Exactly once — each message processed one time.

At least once — messages may be duplicated.

At most once — messages may be lost.

Ordering — messages arriving in sequence.

Causality — cause before effect.

Vector clock — tracking causality across nodes.

Lamport clock — a simple logical clock.

Wall clock — real time.

Clock skew — disagreement between clocks.

Leap second — an extra second in time.

Time zone — a region's offset from UTC.

Daylight saving — seasonal clock changes.

Epoch — a reference point for time.

Timestamp — a point in time.

Duration — a length of time.

Calendar — dates and their rules.

Locale — language and region settings.

Internationalization — preparing for many languages.

Localization — translating for a language.

Unicode — a standard for all characters.

Encoding — bytes for characters.

UTF-8 — a variable-length Unicode encoding.

Normalization — canonical forms of text.

Collation — sorting text correctly.

Case folding — comparing ignoring case.

Regular expression — a pattern for text.

Parser — a tool that reads structured text.

Grammar — rules for valid input.

Lexer — a tool that splits text into tokens.

Token — a meaningful piece of text.

Abstract syntax tree — a tree of parsed code.

Visitor — a pattern for walking trees.

Interpreter — a tool that runs code directly.

Compiler — a tool that translates code.

Optimization — making code faster.

Constant folding — computing constants at compile time.

Dead code — code that never runs.

Register — a fast storage location.

Instruction — one machine operation.

Assembly — human-readable machine code.

Machine code — bytes the processor runs.

Endianness — byte order in memory.

Word — the processor's natural unit.

Alignment — placing data at friendly addresses.

Cache line — the unit of memory caching.

Prefetch — loading memory before it is needed.

Branch prediction — guessing which way code goes.

Pipeline — overlapping instruction stages.

Vectorization — using wide instructions.

SIMD — one instruction on many values.

GPU — a processor for parallel work.

Kernel — a function run on a GPU.

Thread block — a group of GPU threads.

Shared memory — fast memory within a block.

Coalesced access — neighboring threads reading neighboring memory.

Occupancy — how busy the GPU is kept.

Host — the CPU side.

Device — the GPU side.

Transfer — copying between host and device.

Pinned memory — memory the GPU can copy directly.

Stream — a queue of GPU work.

Synchronization — waiting for work to finish.

Profiler — a tool that measures where time is spent.

Roofline model — a limit from compute and memory.

Arithmetic intensity — work per byte moved.

Memory bound — limited by memory speed.

Compute bound — limited by processor speed.

Latency hiding — overlapping waits with work.

Occupancy — how busy the GPU is kept.

Divergence — threads taking different paths.

Reduction — combining many values into one.

Scan — running totals across values.

Histogram — counting values in bins.

Sort — putting data in order.

Merge — combining sorted data.

Partition — splitting data around a pivot.

Stable sort — sort keeping equal items in order.

Comparison sort — sort using only comparisons.

Radix sort — sort by digits.

Counting sort — sort by counting values.

Bucket sort — sort by distributing into bins.

External sort — sorting data too big for memory.

Merge sort — divide and merge sorting.

Quicksort — fast average-case sorting.

Heapsort — sorting with a heap.

Insertion sort — sorting by inserting in order.

Selection sort — sorting by repeatedly picking the smallest.

Bubble sort — sorting by repeatedly swapping neighbors.

Timsort — a hybrid merge and insertion sort.

Introsort — quicksort with a fallback.

Pivot — the element used to split data.

Median — the middle value.

Quantile — a value at a given fraction.

Percentile — a quantile in hundredths.

Order statistic — the k-th smallest value.

Selection — finding the k-th smallest value.

Binary search — finding by halving the range.

Lower bound — the first position not less than a value.

Upper bound — the first position greater than a value.

Interpolation search — guessing where a value is.

Exponential search — doubling to find a range.

Ternary search — splitting into three parts.

Golden section search — finding a maximum by shrinking.

Gradient descent — following the slope downhill.

Learning rate — how big each step is.

Momentum — remembering previous steps.

Adam — an adaptive gradient method.

Loss function — what we want to minimize.

Overfitting — memorizing instead of learning.

Underfitting — failing to learn the pattern.

Regularization — penalizing complexity.

Validation set — data for tuning.

Test set — data for final evaluation.

Cross-validation — rotating the validation set.

Hyperparameter — a setting chosen before training.

Epoch — one pass over the data.

Batch — a group of examples.

Mini-batch — a small batch.

Stochastic — using random samples.

Deterministic — always the same result.

Seed — the start of a random sequence.

Reproducibility — getting the same result again.

Checkpoint — a saved model during training.

Early stopping — stopping when validation worsens.

Learning curve — loss over time.

Confusion matrix — counts of predictions versus truth.

Precision — how many predictions were right.

Recall — how many true cases were found.

F1 score — the balance of precision and recall.

Accuracy — the fraction correct.

ROC curve — true versus false positive rate.

AUC — the area under the ROC curve.

Threshold — the cutoff for a decision.

Calibration — probabilities matching reality.

Baseline — a simple model to beat.

Ablation — removing parts to see what matters.

Feature — an input to the model.

Feature engineering — designing features by hand.

Feature selection — choosing useful features.

Dimensionality reduction — fewer features.

Principal component analysis — projecting to main directions.

Embedding — a learned vector for a thing.

One-hot encoding — a vector with one 1.

Bag of words — counting words ignoring order.

TF-IDF — word importance by rarity.

N-gram — a run of n words.

Tokenization — splitting text into tokens.

Stemming — cutting words to a root.

Lemmatization — mapping words to dictionary forms.

Stop word — a common word often ignored.

Vocabulary — the set of known words.

Out of vocabulary — a word not in the vocabulary.

Language model — a model of word sequences.

Perplexity — how surprised a model is.

Sequence — ordered data.

Recurrent network — a network with memory.

Hidden state — the memory of a recurrent network.

Vanishing gradient — learning signal fading away.

Exploding gradient — learning signal blowing up.

Gradient clipping — capping the gradient size.

Long short-term memory — a recurrent cell with gates.

Gated recurrent unit — a simpler gated cell.

Bidirectional — reading in both directions.

Encoder — the part that reads input.

Decoder — the part that writes output.

Sequence to sequence — input sequence to output sequence.

Attention — looking at relevant parts.

Self-attention — attention within one sequence.

Transformer — a model built on attention.

Positional encoding — telling order to attention.

Multi-head attention — several attentions in parallel.

Feed-forward network — a plain layer after attention.

Residual connection — adding the input back.

Layer normalization — normalizing activations.

Pre-training — learning before the task.

Fine-tuning — adapting a pre-trained model.

Transfer learning — reusing knowledge for a new task.

Zero-shot — doing a task with no examples.

Few-shot — doing a task with a few examples.

Prompt — the input text to a model.

Completion — the model's output text.

Temperature — how random sampling is.

Top-k sampling — sampling from the k best.

Top-p sampling — sampling from the most likely mass.

Beam search — keeping several candidates.

Greedy decoding — always taking the best next word.

Hallucination — confident but wrong output.

Grounding — tying output to real sources.

Retrieval — finding relevant documents.

Retrieval-augmented generation — answering with found documents.

Vector database — a store for embeddings.

Similarity search — finding nearest vectors.

Cosine similarity — angle between vectors.

Euclidean distance — straight-line distance.

Dot product — a measure of alignment.

Norm — the length of a vector.

Unit vector — a vector of length one.

Matrix — a grid of numbers.

Tensor — a multi-dimensional array.

Broadcasting — stretching small arrays to match.

Vectorization — operating on whole arrays.

Automatic differentiation — computing gradients by the chain rule.

Computation graph — the record of operations.

Forward pass — computing the output.

Backward pass — computing the gradients.

Parameter — a learned weight.

Bias — a learned offset.

Activation function — a nonlinear step.

Rectified linear unit — max of zero and the input.

Sigmoid — a squashing to zero to one.

Softmax — turning scores into probabilities.

Dropout — randomly silencing units.

Batch normalization — normalizing each batch.

Weight decay — shrinking weights each step.

Initialization — the starting weights.

Vanishing gradient — learning signal fading away.

Exploding gradient — learning signal blowing up.

Skip connection — adding the input back.

Dense layer — every input to every output.

Convolution — sliding a small filter.

Kernel — the filter in a convolution.

Stride — how far the filter moves.

Padding — extra border around the input.

Pooling — shrinking by summarizing.

Receptive field — what one unit can see.

Feature map — the output of a convolution.

Channel — one slice of a tensor.

Image classification — naming what is in a picture.

Object detection — finding and naming objects.

Segmentation — labeling every pixel.

Bounding box — a rectangle around an object.

Anchor box — a template box for detection.

Non-maximum suppression — removing duplicate boxes.

Intersection over union — overlap of two boxes.

Mean average precision — a detection score.

Data augmentation — making more training data.

Flip — mirroring an image.

Crop — cutting out a piece.

Rotation — turning an image.

Color jitter — changing colors slightly.

Noise — random pixel changes.

Blur — smoothing an image.

Resize — changing the size.

Normalization — scaling pixel values.

Mean subtraction — centering pixel values.

Standard deviation — scaling pixel values.

Channel order — red, green, blue or blue, green, red.

Color space — how colors are described.

Grayscale — one channel of brightness.

Histogram equalization — spreading out contrast.

Edge detection — finding boundaries.

Gradient — change in brightness