Student AI Hub — Foundation Repository Final Audit

1. What this repository is

This repository documents the initial build of the Student AI Hub foundation. It contains the methodology, scripts, source registry, and five locked reference sections that establish the credibility and audit trail for the Student AI Hub. The foundation represents a documented process for creating citation-grounded reference materials from approved sources, serving as both a historical record and a reference for understanding how sources were selected, how content was organized, and how reference sections were written and reviewed.

2. What this repository is not

- **Not the full Student AI Hub**: This repository contains the foundation and initial reference sections, but the Student AI Hub may include additional content created through other means.
- **Not an AI tool or chatbot**: This repository documents a process for creating reference materials. It does not provide conversational interfaces or answer questions in real time.
- **Not a live or ongoing system**: This repository is a record of how the foundation was built. It is not a running application or service that students interact with directly.
- **Not a guarantee of coverage**: The foundation covers five initial topics. It does not claim comprehensive coverage of all AI topics relevant to students.
- **Not automatically updated**: The sections were created as part of the initial foundation and are marked as "locked." They represent the state of information at the time they were created and reviewed.

3. Repository map (human-readable)

The foundation repository is organized into four main directories:

docs/

Documentation and audit materials: - `00_briefing_packet/` — Overview materials, final outputs, source lists, and audit reports - `01_process/` — Process documentation explaining how the foundation was built - `02_architecture/` — Architecture documentation (may be empty or reserved for future documentation) - `03_policies/` — Policy documentation (may be empty or reserved for future documentation) - `04_audits/` — Audit documents including this file - `05_registry_snapshots/` — CSV snapshots of the source registry at different lifecycle stages

content/

Published reference materials: - `sections/` — Five locked reference sections (AI Basics, Using AI for School and Work, How Businesses Are Using AI, AI Tools You Might Use, Rules, Risks, and Ethics of AI) - `pdf/` — PDF generation configuration - `exports/` — Exported content (if present)

data/

Corpus data and indexes: - `chunks/` — Chunked source content with stable IDs - `indexes/` — Search indexes built from chunks - `registry/` — Source registry CSV - `snapshots/` — Source snapshot files (not tracked in git) - `runs/` — Processing run logs

scripts/

Build and maintenance scripts: - `pipeline/` — Ingestion, indexing, and chunking scripts - `search/` — Search and retrieval scripts - `sheets/` — Google Sheets integration scripts - `utils/` — Utility scripts

4. Canonical artifacts produced (what a reviewer should open)

The following artifacts represent the key outputs and documentation that reviewers should examine:

Overview documents

- `foundation/FOUNDATION_OVERVIEW.pdf` — Ultra-brief overview of the foundation's purpose and role
- `foundation/docs/00_briefing_packet/01_overview/FOUNDATION_BRIEFING.pdf` — Comprehensive briefing document summarizing the foundation, build process, and available materials

Process documentation

- `foundation/docs/01_process/PROCESS_OVERVIEW.md` — Non-technical explanation of the build process
- `foundation/docs/01_process/PROCESS_OVERVIEW.pdf` — PDF version of process overview
- `foundation/docs/01_process/WORKFLOW_APPENDIX.md` — Technical workflow details
- `foundation/docs/01_process/WORKFLOW_APPENDIX.pdf` — PDF version of workflow appendix

Published reference sections

Five locked reference sections available as PDFs: - `foundation/docs/00_briefing_packet/02_outputs/pdfs/ai-` - `foundation/docs/00_briefing_packet/02_outputs/pdfs/using-ai-for-school-and-work.pdf` - `foundation/docs/00_briefing_packet/02_outputs/pdfs/how-businesses-use-ai.pdf` - `foundation/docs/00_briefing_packet/02_outputs/pdfs/ai-tools-you-might-use.pdf` - `foundation/docs/00_briefing_packet/02_outputs/pdfs/rules-risks-ethics.pdf`

Source registry and audit materials

- `foundation/docs/05_registry_snapshots/` — CSV snapshots documenting the source lifecycle
  - `legacy_collected_links_v0.csv` — Original 36 collected links
  - `active_ingested_links_v1.csv` — 29 ingested links
  - `active_used_in_final_sections_v1.csv` — 22 links used in final sections
  - `README.md` — Documentation of snapshot meanings
- `foundation/docs/04_audits/REGISTRY_STATUS_MODEL.md` — Explanation of the three-state registry model
- `foundation/docs/00_briefing_packet/04_audit/UNUSED_SOURCES_EXACT_REASONS.md` — Evidence-based report on unused sources
- `foundation/docs/00_briefing_packet/04_audit/URL_MANIFEST.md` — URL manifest with evidence

- `foundation/docs/00_briefing_packet/04_audit/URL_MANIFEST.csv` — CSV version of URL manifest

Additional audit documents

- `foundation/docs/04_audits/REPO_AUDIT_ALIGNMENT.md` — Repository alignment document
- `foundation/docs/04_audits/README_TRUTH_AUDIT.md` — Truth audit documentation

5. Source registry lifecycle (evidence-based)

The foundation registry distinguishes between three states of source URLs, forming a nested hierarchy:

Collected (36 links)

URLs that were present in the registry at the time the foundation corpus v0 was created. These represent the initial set of sources that were selected and approved for use in building the foundation.

Evidence: `foundation/docs/05_registry_snapshots/legacy_collected_links_v0.csv` (36 data rows)

Ingested into corpus (29 links)

URLs that appear in `foundation/data/chunks/chunks.jsonl` as URLs with associated chunked content. These are links that were successfully scraped, processed, and broken into citable chunks.

Evidence: - `foundation/data/chunks/chunks.jsonl` (source of truth) - `foundation/docs/05_registry_snaps` (29 data rows) - `foundation/scripts/sheets/ingested_urls.json` (computed list)

Meaning: The URL was successfully accessed, its content was retrieved, and it was chunked into the corpus. This does not guarantee the content was used in final sections.

Used in final published foundation sections (22 links)

URLs that are traceable from chunk IDs referenced in published foundation section markdown files back to URLs via chunk metadata. These are links whose chunks were actually referenced in the final locked reference sections.

Evidence: - `foundation/content/sections/**/index.md` (chunk ID references) - `foundation/data/chunks/ch` (chunk ID to URL mapping) - `foundation/docs/05_registry_snapshots/active_used_in_final_sections_v1` (22 data rows) - `foundation/scripts/sheets/used_in_final_sections_urls.json` (computed list)

Meaning: The URL's chunks were referenced in at least one published foundation reference section. This is a subset of ingested links.

Subset relationship

The three states form a strict nested hierarchy:

- `active_used_in_final_sections_v1.csv` (22) `active_ingested_links_v1.csv` (29) `legacy_collected_links_v0.csv` (36)

Gaps: - 7 links were collected but not ingested (36 - 29 = 7) - 7 links were ingested but not used in final sections (29 - 22 = 7)

No inference is made about quality, relevance, or intent based on non-use. These gaps are documented but not explained in this audit.

Documentation: See `foundation/docs/04_audits/REGISTRY_STATUS_MODEL.md` and `foundation/docs/05_registry_snapshots/README.md` for detailed explanations.

6. Build pipeline summary (non-technical)

The foundation build process followed these steps:

Step 1: Source registry creation

A Google Sheet was created containing 36 approved URLs, each selected by a human, reviewed for credibility and relevance, assigned to a specific section, and labeled by source type. This registry defines the initial set of sources used to build the foundational reference sections.

Evidence: `foundation/data/registry/SAIH Content - Corpus v0.csv`

Step 2: Controlled ingestion

For each approved URL, the system attempted to fetch and process the content. In this context, 'the system' refers to the documented scripts and workflows in this repository, not an autonomous AI agent. Pages were fetched only if publicly accessible. The system respected robots.txt and site restrictions, did not bypass paywalls or gated content, and recorded failures clearly. When allowed, the system extracted page metadata, headings, and full text. When not allowed, it stored only metadata and marked the source as blocked.

Evidence: `foundation/data/chunks/chunks.jsonl` (successful ingestions), `foundation/data/snapshots/` (snapshot files, not tracked in git)

Step 3: Chunking and indexing

Full texts were split into small, readable chunks (approximately 1200-1800 characters each). Each chunk belongs to exactly one approved source, has a stable ID, and can be cited directly. An index was built to enable search and retrieval of relevant chunks.

Evidence: `foundation/data/chunks/chunks.jsonl`, `foundation/data/indexes/`

Step 4: Content drafting

Using the chunked corpus, draft reference sections were created. The system searched for relevant chunks based on section topics and used those chunks to draft content. All content was grounded in the ingested chunks; no new information was added.

Evidence: `foundation/content/sections/**/index.md` (draft markdown files)

Step 5: Human review and revision

The generated reference sections were reviewed for source balance, over-reliance on single sources, prescriptive or moralizing language, and unsupported claims. When issues were found, language

was narrowed or scoped rather than expanded. No new sources were added during revision. The sections were then finalized and marked as "locked."

Evidence: `foundation/content/sections/**/index.md` (final locked versions)

Step 6: PDF generation

The finalized markdown sections were rendered into PDFs using Pandoc and LaTeX for high-quality typesetting.

Evidence: `foundation/docs/00_briefing_packet/02_outputs/pdfs/*.pdf`

Process documentation: See `foundation/docs/01_process/PROCESS_OVERVIEW.md` and `foundation/docs/01_process/WORKFLOW_APPENDIX.md` for detailed technical and non-technical explanations.

7. Reproducibility notes (bounded)

What can be reproduced locally

The following components can be rerun using scripts in the repository:

- **Chunking**: `foundation/scripts/pipeline/chunk_corpus.py` — Processes snapshots into chunks
- **Indexing**: `foundation/scripts/pipeline/build_corpus_index.py` — Builds search indexes
- **Search**: `foundation/scripts/search/search_chunks_v2.py` — Searches chunks using BM25
- **PDF generation**: Pandoc commands (see existing PDF generation process)

What cannot be fully reproduced

- **Ingestion**: Requires access to source websites and may be blocked by robots.txt or access restrictions. Snapshot files are not tracked in git (`foundation/data/snapshots/` is intentionally excluded).
- **Google Sheets integration**: Requires OAuth credentials (`foundation/scripts/sheets/credentials.js` and `foundation/scripts/sheets/token.json` are not tracked in git).
- **Registry updates**: The Google Sheet registry requires manual access or API credentials.

Intentionally excluded from git

The following data is intentionally not tracked in git: - `foundation/data/snapshots/` — Source snapshot files (large, may contain copyrighted content) - `foundation/scripts/sheets/credentials.json` — Google OAuth credentials - `foundation/scripts/sheets/token.json` — OAuth tokens - `foundation/scripts/sheets/.env` — Environment variables including sheet IDs

This exclusion is documented in `.gitignore` files.

8. Known limits and open questions

The following limits are established by repository evidence:

- **Not all collected links were ingested**: 7 of 36 collected links were not successfully ingested into the corpus. Reasons may include access restrictions, robots.txt blocks, or paywalls, but specific reasons are not documented in this audit.
- **Not all ingested links were used in final sections**: 7 of 29 ingested links were not used in the final published reference sections. See `foundation/docs/00_briefing_packet/04_audit/UNUSED_SOURC` for evidence-based explanations.
- **Human-authored hub pages are outside foundation**: Certain sections (AI News That Matters, Penn State AI Resources, AI by Smeal Major) were intentionally excluded from the automated pipeline and are intended to be human-written. This boundary is documented in process documentation.
- **Foundation covers five initial topics only**: The foundation includes five reference sections. It does not claim comprehensive coverage of all AI topics relevant to students.
- **Snapshots not tracked in git**: Source snapshot files are intentionally excluded from version control, meaning full reproducibility requires re-ingestion or access to snapshot archives.
- **No ongoing updates**: The foundation sections are marked as "locked" and represent a point-in-time snapshot. They are not automatically updated.

9. Audit trail index

The following audit documents exist in the foundation repository:

Registry and source audits

- `foundation/docs/04_audits/REGISTRY_STATUS_MODEL.md` — Explanation of the three-state registry model (collected, ingested, used)
- `foundation/docs/05_registry_snapshots/README.md` — Documentation of registry snapshot artifacts
- `foundation/docs/00_briefing_packet/04_audit/UNUSED_SOURCES_EXACT_REASONS.md` — Evidence-based report explaining why certain approved sources were not used in final sections
- `foundation/docs/00_briefing_packet/04_audit/URL_MANIFEST.md` — URL manifest distinguishing approved, ingested, and used URLs with evidence
- `foundation/docs/00_briefing_packet/04_audit/URL_MANIFEST.csv` — CSV version of URL manifest

Repository audits

- `foundation/docs/04_audits/REPO_AUDIT_ALIGNMENT.md` — Repository alignment document describing purpose, features, artifacts, and gaps
- `foundation/docs/04_audits/README_TRUTH_AUDIT.md` — Truth audit documentation
- `foundation/docs/04_audits/FOUNDATION_REPOSITORY_FINAL_AUDIT.md` — This document

Process documentation

- `foundation/docs/01_process/PROCESS_OVERVIEW.md` — Non-technical process overview
- `foundation/docs/01_process/WORKFLOW_APPENDIX.md` — Technical workflow details

Scope and boundaries

- `foundation/docs/FOUNDATION_SCOPE.md` — Authoritative positioning document defining the repository's role and scope

---

Document version: Final audit v1
Date: 2026-01-22
Audit scope: Foundation repository (`foundation/` directory)
Evidence basis: Repository artifacts only, no external claims