

Student AI Hub — Process Overview

Student AI Hub — Process Overview

This document explains how the initial foundational framework and core reference content of the Student AI Hub were built—from a spreadsheet of approved links to a set of citation-grounded reference sections.

Scope and Purpose

Important: This process describes how the hub's foundation, structure, and initial canonical reference sections were established at the outset. The Student AI Hub may also include human-authored, editorial, or institution-specific content that does not flow through this pipeline. The source registry and ingestion process define a trusted reference layer for the hub's core sections, not a restriction on future or human-authored content.

The Process

1. Human-Approved Source Registry

The initial foundation is a Google Sheet that serves as a source registry. Each row represents one approved source that was chosen by a human, reviewed for credibility and relevance, assigned to a specific section, and labeled by source type. Nothing appears in the hub's initial reference sections unless it first exists in this registry.

2. Controlled Ingestion

For each approved URL, the system:

- Fetched pages only if publicly accessible
- Respected robots.txt and site restrictions
- Did not bypass paywalls or gated content
- 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. Nothing was guessed or filled in.

3. Auditable Source Records

Each source was stored as a structured snapshot file including where the content came from, when it was retrieved, whether it was fully accessible, and exactly what text was available. This means every summary can be traced back to a specific source and retrieval moment.

4. Citable Chunks

Full texts were split into small, readable chunks. Each chunk belongs to exactly one approved source, has a stable ID, and can be cited directly. This allows the system to reference evidence precisely instead of loosely summarizing entire articles.

5. AI-Assisted Synthesis with Boundaries

AI tools were not allowed to browse the web or add new information. They were used only to summarize existing chunks and draft section content using those chunks. All AI output was treated as a draft, not an authority.

6. Human Audit and Revision

Each generated reference section was 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.

What's Intentionally Not Automated

Certain sections were not generated from the corpus and are intended to be human-written due to time sensitivity, institutional ownership, or curricular nuance:

- AI News That Matters - Penn State AI Resources
- AI by Smeal Major

This boundary is explicit and documented.

What We Have Now

The following five initial reference sections were synthesized from the corpus and locked:

- **AI Basics** — Core concepts and definitions explaining how AI and machine learning work
- **Using AI for School and Work** — Practical, responsible guidance for students using AI in coursework and productivity
- **How Businesses Are Using AI** — High-level explanation of current business applications of AI
- **AI Tools You Might Use** — Categories of AI tools students may encounter, without endorsement
- **Rules, Risks, and Ethics of AI** — Ethical, legal, and governance considerations surrounding AI

Each section is grounded in human-approved sources, avoids unsupported claims, and can be traced back to specific source snapshots and chunks.

The Result

The Student AI Hub's initial foundation and core reference sections:

- Are grounded in human-approved sources
- Avoid hallucination and silent assumptions
- Are transparent and auditable
- Can be updated without rewriting everything

AI assisted the process, but humans retained control at every decision point. The hub's structure and reference layer established through this process provide a foundation for both automated and human-authored content going forward.