

AI Archive Search Example for Universities and Newspaper Archives

Lehigh's The Brown and White made 400 million words of archive searchable

Lehigh University student Nina Cialone used CustomGPT.ai to build a no-code AI archive search for The Brown and White, the university's student newspaper. More than a century of journalism is now queryable through conversational AI, with cited answers, built by a student in one semester.

400M+

Words indexed

1,400+

Data formats supported

0

Lines of code written

THE CHALLENGE

More than a century of journalism, too big to search by hand

Faculty mentor Craig Gordon challenged Nina to build an AI agent trained on the entire archive of The Brown and White, a publication that dates back to the 19th century.

The agent had to support journalistic research, give structured access to articles in multiple formats and eventually include podcasts and multimedia. Nina and Craig weren't sure any solution could handle an archive that size.

WHAT THE PROJECT NEEDED**Scale.**

Hundreds of millions of words, ingested without custom infrastructure.

Verifiable answers.

Cited facts that journalism can rely on.

No code.

A platform a student could configure alone.

Room to grow.

Support for podcasts and multimedia.

HOW THEY SOLVED IT

A no-code AI agent trained on the full archive

Sitemap ingestion.

Sitemap generation and crawling tools pulled in the archive, replacing hours of copying and pasting.

Every format indexed.

Content was organized across formats, with more than 1,400 supported.

Tuned with beta testers.

Nina customized the persona and refined accuracy with editors and advisors.

Deployed in Slack.

The agent was prepared for editorial use inside the newsroom's existing tool.

ABOUT THE BROWN AND WHITE

The Brown and White is the student newspaper of Lehigh University in Bethlehem, Pennsylvania, publishing since the 19th century. Nina Cialone, a senior studying cognitive science, writes for the paper and covers AI for the Substack publication Don't Count Us Out Yet.

AT A GLANCE

ORGANIZATION	BUILDER	DEPLOYMENT
The Brown and White, Lehigh University's student newspaper	Nina Cialone, student, with faculty mentor Craig Gordon	No-code AI agent, prepared for Slack

HOW CAN AI PREVENT FABRICATED HISTORICAL FACTS?

RAG grounding, source citations and conversational archive search

- 1 Grounded in real articles.

The AI retrieves from the indexed archive before answering, rather than from general training data.
- 2 Citations on every answer.

Users can check each fact against the original article.
- 3 Declines when unsure.

If the archive doesn't hold a reliable answer, the system says so instead of guessing.

RESULTS

Why The Brown and White chose CustomGPT.ai

CRITERION	WHY IT MATTERED	HOW CUSTOMGPT.AI DELIVERED
Sitemap ingestion	Archive spread across many URLs	Automated crawl replaced manual copying
Scale	400M+ words exceeded most no-code tools	Full corpus, no custom infrastructure
Anti-hallucination	Journalism needs verifiable facts	RAG grounding with source citations
No-code	Student builder, no engineering	Full configuration in the UI
1,400+ formats	Podcast and multimedia plans	Audio, video and document support

Nina indexed more than 400 million words, began adding podcast episodes and other multimedia, and beta tested the agent with editors and advisors before preparing it for editorial use in Slack.

//

The specific tools to help create a sitemap were immensely helpful for us because of the way that our archive is set up. Instead of many hours of copying and pasting, all I had to do was just copy and paste the whole thing right into CustomGPT.ai's tool.

Nina Cialone

STUDENT WRITER, THE BROWN AND WHITE

WHY IT WORKED

- Sitemap tools.

The whole archive, ingested fast.
- Format support.

Text today, multimedia next.
- No-code interface.

Focus on content, not code.
- Persona control.

Tuned for each type of reader.

ABOUT CUSTOMGPT.AI

CustomGPT.ai is a no-code platform for building AI agents that answer only from an organization's own content: documents, websites and knowledge bases. Every answer is grounded in verified sources with citations, and the platform is SOC 2 Type 2 and GDPR compliant. Organizations use it for customer support, internal knowledge search, website chat and more, across help centers, apps and the API.