
EduFire - Personalized Education platform using LLM steering

Atir Petkar
State Street

With
Goodfire and Apart Research

Abstract

EduFire is a personalized education platform designed to tailor educational content and assessments based on individual user preferences. Utilizing the Goodfire API for AI model steering, EduFire customizes learning experiences by adjusting Large Language Model (LLM) outputs according to user-selected features. The platform integrates Flask for the backend, FAISS for vector storage, and OpenAI embeddings for document indexing and retrieval. Users are onboarded through preference selection, which informs the model's output in terms of content format and assessment style. The platform allows users to upload documents on specific topics, which are then processed to generate personalized coursework and assessments. The methodology involves steering the LLM's behavior using Goodfire's model variants and features, enabling a high degree of personalization. Results indicate that EduFire successfully generates customized educational materials and assessments, enhancing user engagement and learning efficacy. The project highlights significant implications for AI interpretability, demonstrating how model steering can be applied in educational contexts to meet individual learner needs. Future work includes expanding assessment types, improving the frontend interface, and exploring the scalability of the platform.

Keywords: AI Observability, Mechanistic Interpretability, Model Reprogramming

1. Introduction and Problem Statement

The rapid advancement of AI and machine learning technologies has opened new avenues for personalized education. Traditional educational platforms often lack the flexibility to adapt to individual learner preferences, leading to suboptimal learning experiences. The challenge lies in creating a system that can dynamically adjust content and assessment styles to meet the unique needs of each user.

EduFire addresses this gap by leveraging the Goodfire API for AI model steering, enabling the customization of educational content based on user preferences. The platform aims to enhance learner engagement and efficacy by providing tailored coursework and assessments. This paper discusses the development and implementation of EduFire, focusing on its methodology, results, and implications for AI interpretability.

2. Methodology and Approach

Technologies Used: Backend with Flask (Python), Vector Store with FAISS, Embedding Model with OpenAI Embeddings, LLM for Content Generation with Goodfire API with Model Variants, Frontend with React, TypeScript, Tailwind CSS, shaden/ui, Vite

User Onboarding

Users begin by selecting their content format and assessment style preferences:

- **Content Format Preferences:**
 - Detailed comprehensive notes
 - Concise bullet points
 - Real-world examples
 - Step-by-step guides
- **Assessment Style Preferences:**
 - Multiple-choice quizzes
 - Open-ended questions

These preferences are mapped to corresponding Goodfire features, and the model variant is updated accordingly using feature values.

Experimentation with Goodfire Model Variants

A crucial part of EduFire’s development involved experimenting with Goodfire’s model variants to understand how adjusting feature values affects the AI’s output. The steps included:

- **Feature Retrieval and Mapping:**
 - User preferences for content preferences and assessment preferences are collected during onboarding (e.g., preferred content format and assessment style) were mapped to specific features within the Goodfire model.
 - Relevant features were retrieved using descriptive texts corresponding to user selections.
- **Adjusting Feature Values:**
 - Tested different feature values ranging from -1 to 1 to observe changes in the model’s behavior.
 - Positive values steered the model toward a feature (e.g., setting a feature value to 0.75 to emphasize detailed notes).
 - Negative values steered the model away from a feature.
- **Observing Model Output:**
 - Generated sample outputs after each adjustment to assess the impact on content and assessment generation.
 - Noted how the level of detail, style, and format of the educational materials changed in response to different feature values.

Retrieval Augmented Generation (RAG)

Document Upload and Indexing

- Users/Admin upload a document on a specific topic.
- The document is chunked into manageable pieces.
- Embeddings for each chunk are generated using OpenAI embeddings.
- Embeddings are stored in a FAISS vector store.

Coursework Creation

- The Goodfire model variant generates course notes from the uploaded document, tailored to the user’s content preferences.

- Notes are presented in the selected format (e.g., bullet points or step-by-step guides).

Conversational Q&A

- Users interact with the platform through a chat interface.
- Queries are matched with relevant document chunks retrieved from the FAISS index.
- Responses are generated using the Goodfire model variant, considering both the retrieved content and the conversation history.

Assessment Generation

Assessments are created based on the user's selected assessment style:

- The Goodfire model variant generates assessment items (e.g., multiple-choice questions or open-ended prompts).
- Assessments are delivered to the user in the preferred format.

3. Results and Analysis

EduFire successfully demonstrates the feasibility of personalized education through AI model steering:

- **Personalization:** Users receive customized content and assessments aligned with their preferences.
- **User Engagement:** Tailored materials enhance user engagement by providing relevant and interesting content.
- **Content Quality:** The Goodfire model variant produces coherent and contextually appropriate educational materials.
- **Efficiency:** The integration of FAISS and OpenAI embeddings allows for efficient retrieval of relevant content during Q&A sessions.

4. Discussion of Implications for AI Interpretability

EduFire showcases the practical application of AI interpretability and model steering in education:

- **Model Steering:** Adjusting feature values in the Goodfire model variant demonstrates how AI behavior can be guided to meet specific objectives.
- **Transparency:** Users indirectly influence the AI's output, promoting transparency in how content is generated.
- **Customization vs. Generalization:** The platform balances the need for personalized content with the generalization capabilities of LLMs.

These implications suggest that AI interpretability techniques can be effectively employed to enhance personalized learning, offering insights into how AI models can be tailored without compromising their foundational capabilities.

5. Conclusion and Future Work

EduFire marks a significant advancement in personalized education by integrating AI model steering with user preferences to customize educational content and assessments, thereby enhancing user engagement and learning outcomes.

Future Work:

- **Enhanced User Onboarding:** Expand the onboarding process to capture more detailed learner preferences—such as spaced repetition techniques, preferred sensory modalities (visual or auditory), optimal study times, and learning pace—to better tailor the learning experience to individual needs.
- **Diverse Assessment Methods:** Develop a broader range of assessment types, including project-based assignments, peer assessments, and adaptive quizzes that adjust to the learner's proficiency level, providing more meaningful evaluations.
- **Improved Coursework Personalization:** Implement advanced features like content pacing, multimedia integration, adaptive learning paths, and spaced repetition to reinforce knowledge retention and accommodate different learning styles.
- **Addressing Current Limitations:** Focus on resolving existing issues to enhance system stability, improve user experience, and ensure content accuracy by optimizing the AI model's responsiveness and refining system integration.
- **Frontend Development:** Enhance the user interface to create a more intuitive and engaging experience, accommodating the expanded onboarding process and presenting content and assessments accessibly.
- **Scalability and Performance:** Explore scaling methods to support a larger user base and multiple concurrent sessions, ensuring consistent performance as the platform grows.
- **Advanced Analytics and Adaptation:** Integrate analytics to monitor user progress and engagement, enabling the system to adapt content and assessments dynamically for further personalization and to address individual learner challenges.

By developing EduFire along these lines, we aim to contribute to the evolution of personalized education and AI interpretability. The platform's ongoing enhancements will focus on delivering a highly customized learning environment that adapts to individual preferences, ultimately improving learning outcomes and user satisfaction.

6. Appendix

1. Goodfire. (2024). Goodfire API. Retrieved November 24, 2024, from <https://docs.goodfire.ai/>
2. **FAISS: A library for efficient similarity search.** Facebook AI Research. Retrieved from [FAISS Library](#)
3. **OpenAI Embeddings.** OpenAI. Retrieved from [OpenAI Embeddings](#)