For this project, I propose using a dataset containing articles and research papers related to various fields of artificial intelligence applications. The dataset will include articles covering topics such as natural language processing, computer vision, machine learning algorithms, robotics, and other AI-related areas.

I choose this dataset because it provides a comprehensive and diverse collection of information relevant to the field of artificial intelligence. By leveraging this dataset, the AI model can learn from a wide range of AI applications and research findings, allowing it to provide accurate and informative answers to user queries.

**2. Methodology:**

a. Data Preprocessing: The dataset will undergo preprocessing to extract the most useful information, such as the article text, title, keywords, and abstract. Text cleaning techniques, such as removing stopwords, punctuation, and special characters, will be applied. Additionally, text normalization methods, such as stemming or lemmatization, may be used to reduce variation in word forms. The dataset will then be tokenized and vectorized to prepare it for input into the machine learning model.

b. Machine learning model: The goal of this project is to develop a question-answering system that can provide accurate answers to user queries based on the information contained in the dataset. To achieve this, I propose using a BERT-based question-answering model. BERT (Bidirectional Encoder Representations from Transformers) has shown remarkable performance in natural language understanding tasks, including question answering. By fine-tuning BERT on the dataset, the model can learn to extract relevant information from the articles and generate accurate answers to user queries.

Alternative models considered: Other alternative models that could be considered include traditional machine learning algorithms such as TF-IDF (Term Frequency-Inverse Document Frequency) or deep learning architectures such as LSTM (Long Short-Term Memory) networks. However, BERT is chosen for its ability to capture contextual information and handle complex language understanding tasks effectively.

c. Evaluation Metric: The performance of the question-answering model will be evaluated using metrics such as accuracy, precision, recall, and F1-score. Additionally, confusion matrices may be used to analyze the model's performance in more detail, particularly in cases of incorrect predictions. The baseline accuracy will be established based on simple keyword matching or TF-IDF approaches and will serve as a benchmark for comparison.

**3. Application:**

The user inputs a question or query related to artificial intelligence applications through a simple text input field on a web application interface. The user can also browse through categories or topics to refine their search.

The output displayed to the user is the most relevant and accurate answer to their query, extracted from the articles in the dataset.