

## Artifimo - a Unified and Ultra-Powerful AI Platform

Martin Kostov<sup>1\*</sup>, Presian Tsvetkov<sup>1</sup>, Fabien Kunis<sup>1,2</sup>

<sup>1</sup>125 Boyan Penev High School, Sofia, Bulgaria

<sup>2</sup>Sofia University "St. Kliment Ohridski", Sofia, Bulgaria

Received 21 September 2024, Accepted 21 October 2024

DOI: 10.59957/see.v9.i1.2024.5

---

### ABSTRACT

*According to IBM, 34 % of companies are using AI in one way or another, and this percentage is expected to rise further and further [1]. A major obstacle for companies and individual users is the difficulty in selecting and using appropriate AI models for specific tasks. That's why we created Artifimo, a modern, easy-to-use web application that allows users to have conversations with a wide range of AI models with different skill sets. These obstacles often arise due to the overwhelming variety of models, each with distinct strengths, and the lack of clear guidance on how to leverage them effectively. With Artifimo, users can test the capabilities of different AI models, each offering different features and specializations. Each model has a different purpose-some specialize in solving math problems, others in creative writing, and still others in writing code. The platform provides a seamless and intuitive interface for users to interact with AI models, making it accessible to both beginners and experienced users.*

*Keywords: AI models, generative AI, fine-tuning, machine learning (ML), inference.*

---

### INTRODUCTION

Artifimo is a powerful and easy-to-use platform that allows users to interact with a variety of AI models in an interactive way. With its intuitive interface, customizable settings, and real-time chat capabilities, Artifimo offers a unique and immersive experience for users to explore the world of AI chatbots. A common challenge for companies using AI is finding models that align with their specific needs while ensuring ease of integration. These challenges include high costs due to using a model too big

for the task, lack of domain-specific models, and limited expertise in fine-tuning AI solutions. Artifimo addresses these issues by providing an accessible and flexible platform that allows users to easily select, test, and fine-tune multiple models.

The platform offers a rich set of features including multiple AI models, real-time chat, easy model selection, flexible settings, message history and customizable design. Additionally, Artifimo provides security through user authentication and authorization features, as well as seamless

---

\*Correspondence to: Martin Kostov, 125 Boyan Penev High School, Sofia, Bulgaria, E-mail: [contact@martinkostov.me](mailto:contact@martinkostov.me)

integration with external APIs to dynamically update and add new models.

Artifimo is built with state-of-the-art technologies such as Next.js for front-end, Radix UI [2] for UI components, Zustand [3] for state management, and Pocketbase [4] for back-end functionality. The platform is hosted on Vercel [5], which provides excellent performance.

Regardless of the technical level of the user looking to learn more about AI, or a developer looking to test and compare different models, Artifimo provides a comprehensive and accessible platform for all needs. With its powerful capabilities and ease of use, Artifimo is the ideal tool for exploring with the latest advances in generative AI.

## EXPERIMENTAL

To get started using Artifimo, follow these steps:

1. Register: create a new account on the Artifimo platform by visiting <https://app.artifimo.com> and providing the necessary information such as your name, email address and password. There is also an option to log in via Google [12].

2. Confirm email address: log in to your email inbox and click the link provided.

3. Model selection: After logging in, you will be presented with a list of available AI models. Browse the models and select the one you want to work with. Explore their capabilities, strengths and weaknesses and depending on what you want to achieve choose a suitable model.

4. Chat Interface: After selecting a pattern, you will be redirected to the chat interface. Here you can start communicating with the selected AI model by typing your messages in the input field and pressing enter or clicking the send button. You can attach files or change the options and configuration for a different result.

5. Conversation History: the chat interface displays the history of conversations between you and the AI model. You can scroll through previous messages to view the context of the conversation.

6. Switch Models: To switch to a different AI model, use the model selection drop-down menu located in the top bar of the app. Select the desired model from the list and the chat interface will update.

7. App Settings: You have the option to change the app options by clicking on the “Settings” button in the top right corner.

## Technologies used

Artifimo is built using the following technologies:

- Front-end: The front-end part of Artifimo is developed using Next.js, a popular React library for building server-side applications. We use TypeScript in tandem with Next.js [13].

- UI Components: the Artifimo UI was created using Radix UI, a collection of non-stylized, accessible and customizable UI components [2].

- Artifimo uses Zustand, a library for managing React state [3].

- API Integration: The platform integrates with an external API to retrieve information about available AI models and their configurations. The API integration is handled using the fetch function, which allows HTTP requests to be made to the server.

- Backend: Artifimo uses Pocketbase as a backend. User settings, chats, patterns and any files are stored there [4].

- Hosting: the application “lives” thanks to Vercel [5].

## Interface

User Interface for Fine-tuning AI Large Language Models from the Artifimo catalog is shown on Fig. 1. The user is able to customize various specific options on how the model behaves and upload their files.

User Interface for Text to Image generation in Artifimo is shown on Fig. 2. There are options for changing the image height and width, strength of the prompt, number of images and model selection.

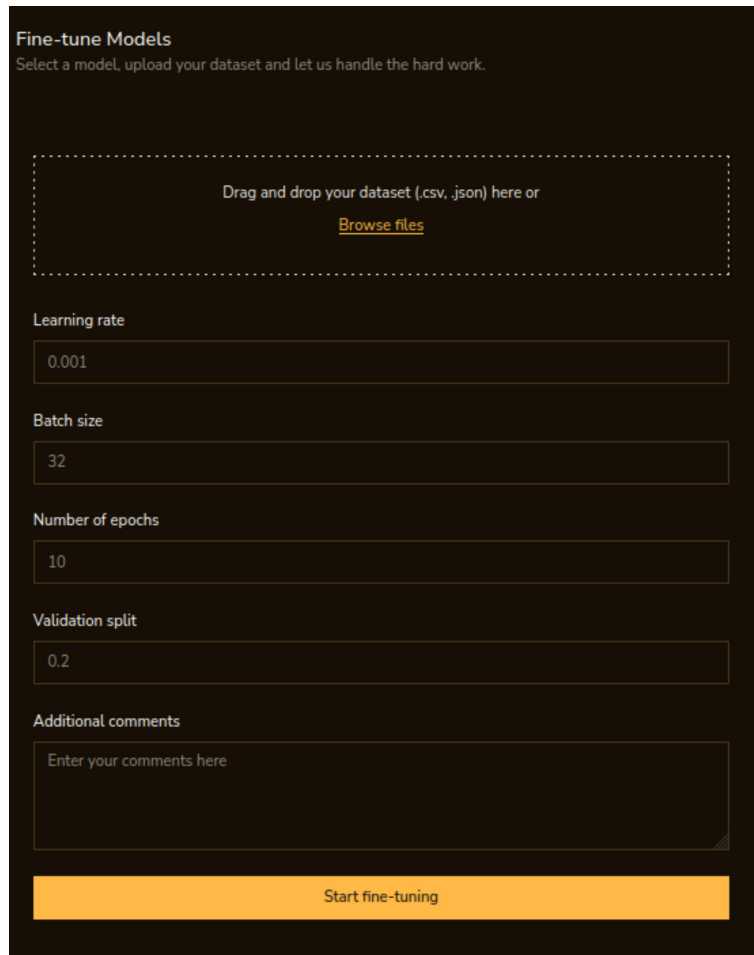


Fig. 1. Fine-tune UI.

The main screen interface for chatting with AI large language models can be seen on Fig. 3. There is a chat history window on the left side of the screen and the main content in the center, along with model selection at the top of the screen and model settings.

## RESULTS AND DISCUSSION

### *Specialized models*

The main goal of the project is to provide users with a powerful and intuitive tool to use different AI models specifically designed for specific tasks. This often eliminates repetition, “hallucinations”, and false information that AI models sometimes present because of the data used to train them.

The benefits of using a different model for each task are numerous, such as:

- Lower cost: if the model is specialized for a given task, it does not need multiple parameters. Instead of a model of 70 billion parameters, a model of 7 billion can be used. This dramatically reduces the cost of using it.
- Faster answers: when the model has fewer parameters it can generate the tokens of each answer much faster.
- Fewer “hallucinations”: when all these factors are combined, the model has less but more correct and precise information that is less likely to confuse it.

The platform aims to facilitate access to these

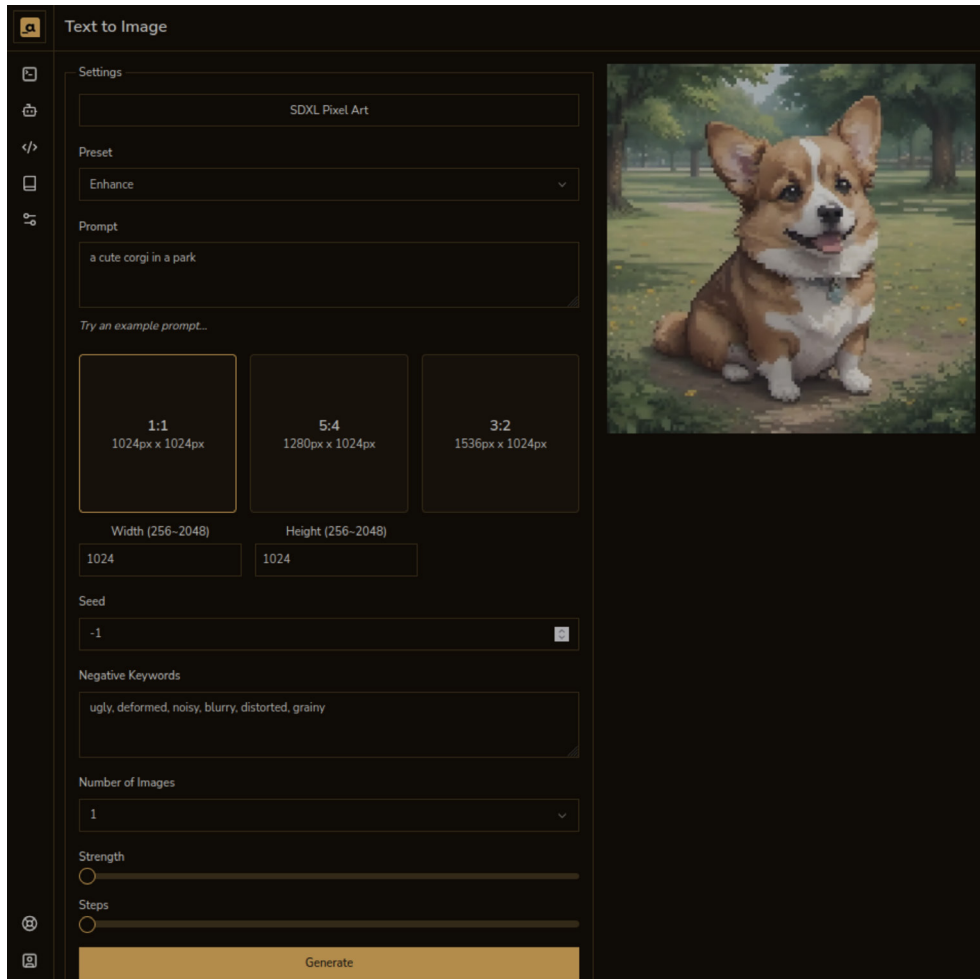


Fig. 2. Text to Image.

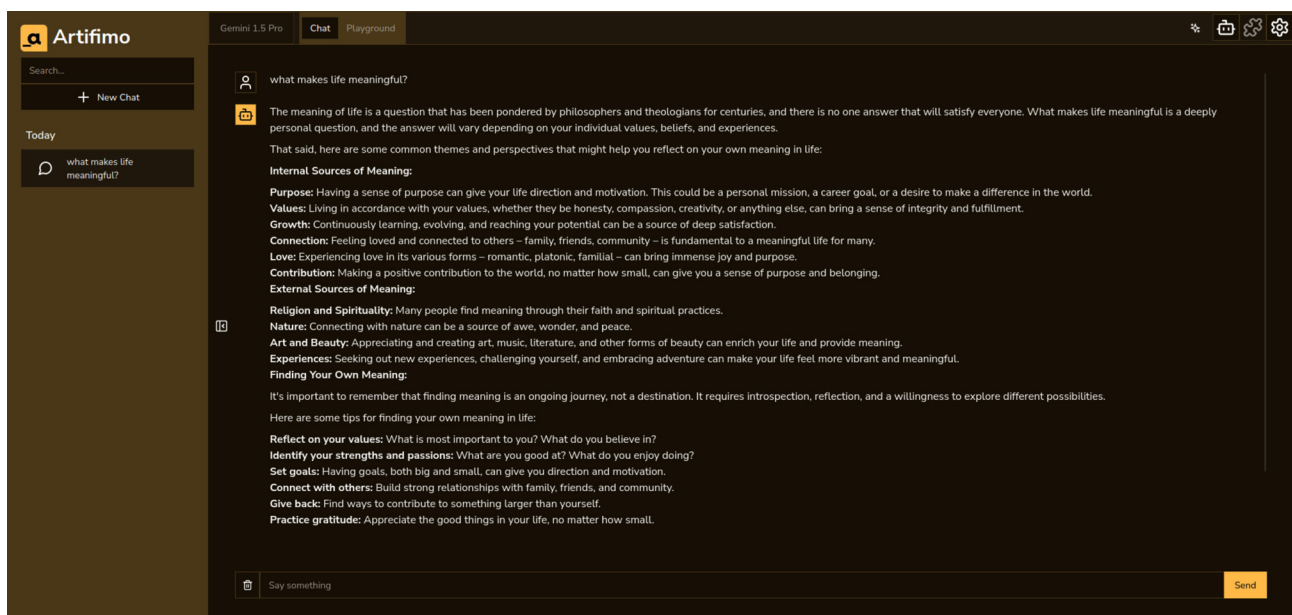


Fig. 3. Chat UI.

models through an intuitive and easy-to-use interface, allowing users with different technical backgrounds to take advantage of the capabilities of generative AI.

### ***Security and data protection***

Along with providing access to a variety of AI models, we place a strong emphasis on data security and privacy. Understanding the sensitive nature of information shared during AI conversations, the platform is committed to implementing robust data protection measures. This includes the use of secure encryption protocols (SSL, HTTPS) to ensure that communication between users and AI models remains confidential and only for them. In addition, Artifimo follows strict data storage and management practices, ensuring that user information is protected from unauthorized access or misuse. One of these practices is an end-to-end encryption method for Pocketbase (the database) [4]. By prioritizing security and privacy, Artifimo aims to build trust among users and provide a trusted and safe environment for AI conversation. We aim to provide it so users can converse on any topic, however personal, with the AI. In addition, we will offer additional data security measures to businesses. One of them is the option of self-hosting on your own server.

### ***Model programming (Fine-tuning)***

Another key goal of the Artifimo project is to offer opportunities for fine-tuning AI models according to the specific needs and preferences of users or businesses. The platform recognises that each user or organization has special requirements and that a one-size-fits-all approach is not suitable for all. We develop a user-friendly UI where users can attach a document with their own data (.csv, .json for example) and using our server space, Artifimo tunes the model according to the requirements and data. Once this operation is complete there is an API service to use the model [3] - again on our servers (aka “inference”). In this way, the model “learns” information it did

not have in mind before this process.

### ***Corporate Client Services***

We know that companies often handle sensitive documents and company data that require the highest level of security and confidentiality. Artifimo strives to be a trusted partner for organizations looking to leverage generative AI for their own, internal use. The platform aims to provide businesses with a secure environment in which to upload and manage their documents without worrying about unauthorized access or privacy breaches.

Moreover, Artifimo understands that organizations may have specific requirements and need customized AI models tailored to their specific use cases. In addition to the previously mentioned model training feature, we offer the ability for businesses to upload their own models trained on their data, enabling seamless integration with their existing systems and workflows.

In realizing these goals through Artifimo, we aim to be a leader in the field of generative artificial intelligence.

### ***Stages of implementation***

We divide the project into three main stages of realization: Test stage, First version, Finished product.

#### ***Test stage***

This stage includes all the new features we are developing. After a long period of testing and making sure that a new feature works as we’ve designed it, it goes into the “First Release” stage. At the time of writing the following features are in the testing stage:

- Training and serving a model
- Attaching files
- Browsing and internet access
- Access for corporate users

Even though some features are in the testing stage some of them can still be used but with limited access.

### First version

At the time of writing Artifimo is in this phase of development. The app is available for testing currently, and we invite all users to try it out and provide us with feedback on the good and the bad.

### Final product

This is the final stage of implementation. When Artifimo reaches this stage, the application will be available to everyone, and all features will be complete and fully ready for use. The app will also be available on Google Play and Apple App Store for mobile devices.

### Logical description

Fig. 4 represents a sequence of interactions between the software architecture of Artifimo, designed to provide AI-generated responses to user queries. The process begins with the user interacting with the user interface on the front-end, usually by sending a request to the server by typing a query into an AI model and pressing

the enter key on the keyboard. The front-end then sends the user's request to the back-end, along with the message details, the chat history and (if any) customizations or instructions that the user has made to the AI model. The back-end authenticates the request and the user via PocketBase, a backend-as-a-service that manages authentication and storage. It retrieves previous chats, memory and context and saves the new one. Once authenticated, the back-end sends a request to the AI models for a response from one of the available providers - Hosted (ready-to-use models that are provided by Artifimo) or Local (custom models that the user can upload from their machine with WebLLM [6]). The AI models generate the response and save the conversation data back to PocketBase. PocketBase then sends the AI response to the back-end, which forwards it to the front-end. Finally, the front-end displays the AI response to the user. This sequence runs and completes within a few seconds and ensures that the requests are securely authenticated,

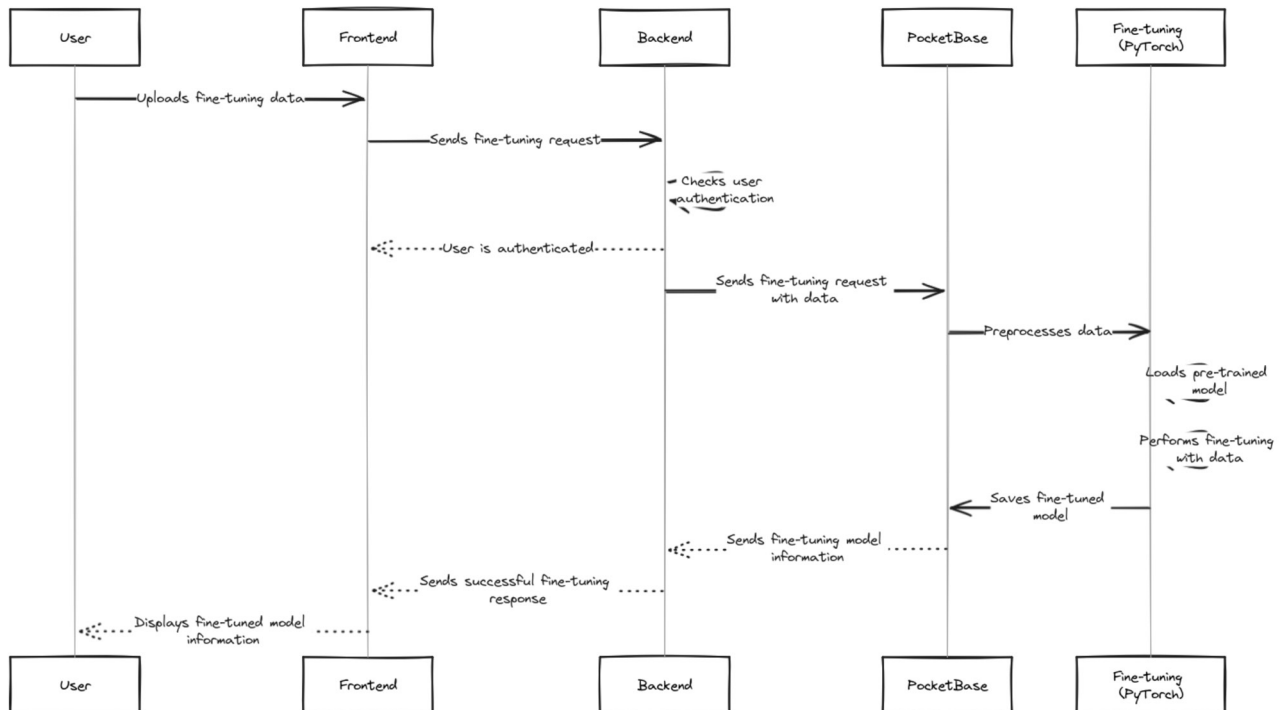


Fig. 4. Conversation process with generative model.



processed, and responded to by leveraging the encryption capabilities with PocketBase.

Fig. 5 illustrates the sequence of functions within the Artifimo fine-tuning system, designed to handle teaching AI models with new data that has been previously outside their training dataset and therefore unknown to them. The process begins when the user uploads the data via the front-end interface. The front-end then sends this request, along with the data, to the back-end. The back-end checks user authentication by communicating with PocketBase to validate the user's credentials, same as the process in Fig. 1. Upon successful authentication, the back-end forwards the request and data to PocketBase where the files will be saved for future reference. After a successful upload to PocketBase, the data is processed to ensure it is suitable for the task. If it is, the request proceeds, and if not, the user is presented with an error asking them to upload new data along with details on what went wrong.

Next, the data from PocketBase gets sent to a server where the model will be fine-tuned

with PyTorch [7]. The server then begins the fine-tuning process using the provided data. It loads a pre-trained model and its tokenizer from HuggingFace [8]. The dataset is prepared by tokenizing the text, ensuring it fits the model's input requirements. Data loaders are created to batch and shuffle the data for efficient processing. The model is trained by running multiple epochs, using the AdamW optimizer [9], which includes weight decay for better generalization. The training process minimizes the cross-entropy loss function (a popular loss function used in machine learning to measure the performance of a classification model), which measures the difference between the predicted output and the true labels. Throughout the training, the model's weights are updated to reduce this loss. Finally, to evaluate the model's performance, a test set is used to measure the accuracy, the precision, and other relevant metrics. Once complete, PocketBase saves the updated model details and sends the updated model information back to the back-end. The back-end confirms the

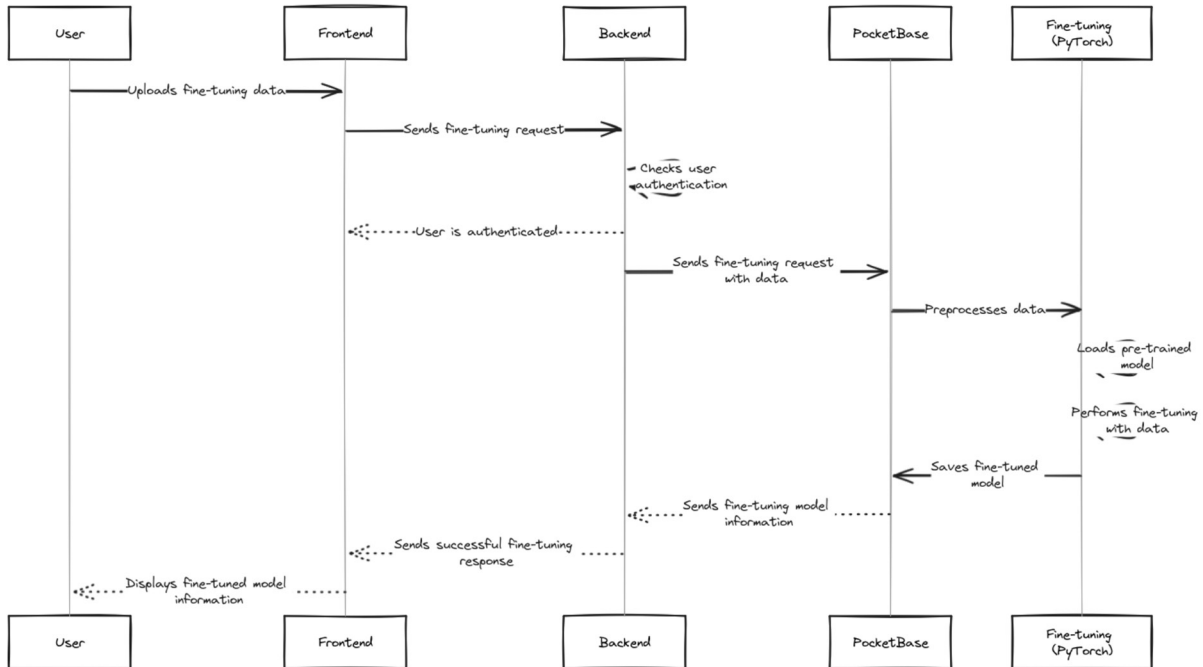


Fig. 5. Training and teaching a model.

successful task and sends a response to the front-end, containing the model path (the unique file name for the model) and it appears in the model selector. Finally, the front-end displays the model to the user, completing the process.

Fig. 6 represents the PocketBase database schema for Artifimo, which is designed for managing users, both corporate and individual, along with their subscriptions, resources, and

interactions with AI models. The main tables include User, CorporateUser, IndividualUser, Subscription, CorporateResources, SubscriptionModel, Parameters, and Resources. The User table holds general information like name, email, password, and profile picture, and methods for authentication and profile updates. CorporateUser and IndividualUser extend this to include specific details relevant to each type,

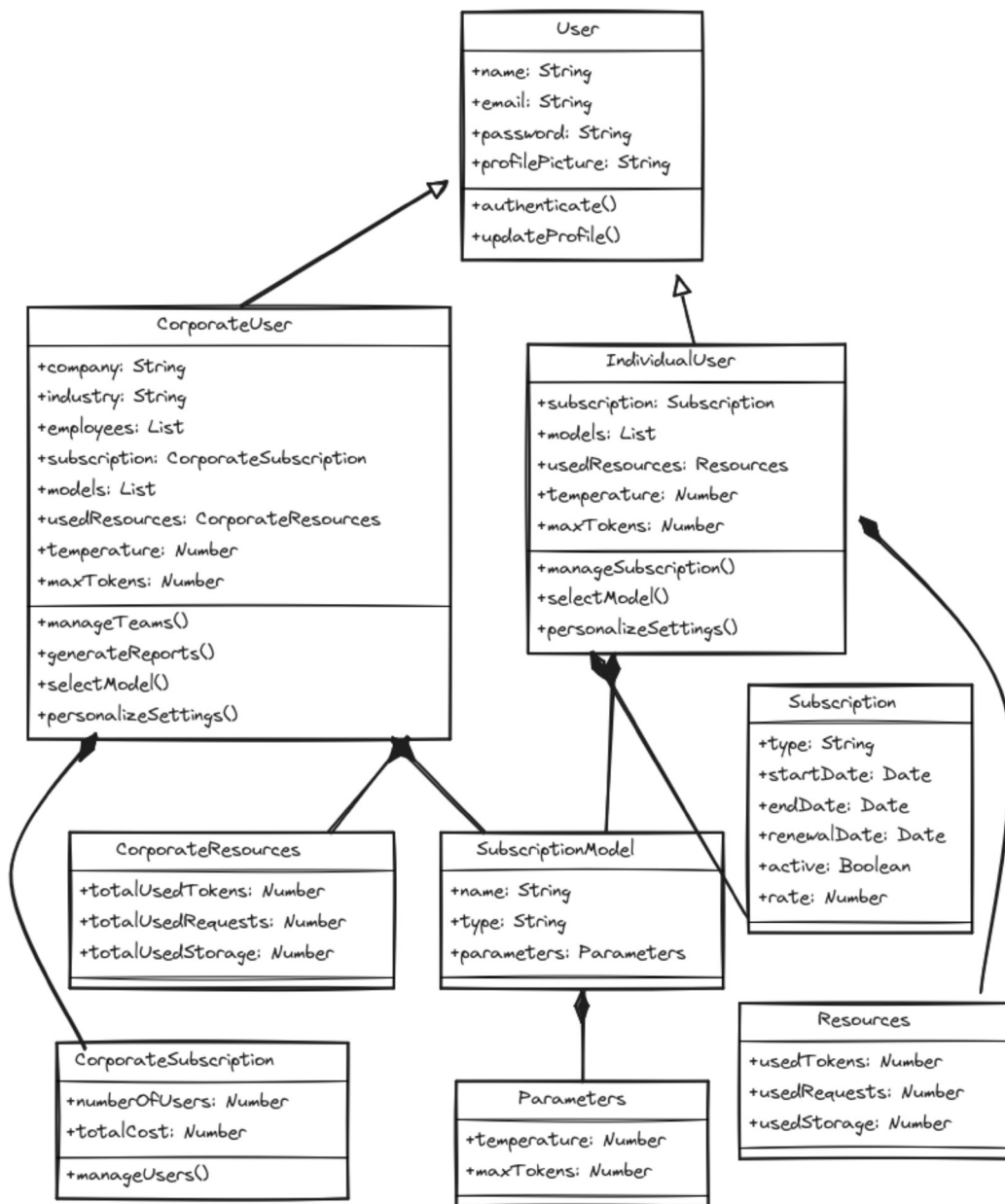


Fig. 6. Database schema.



such as company information for corporate users and subscription details for individual users. The subscription is available for users and companies who wish not to handle the fine-tuning process themselves. Otherwise, they can go through the process in the UI, represented by Fig. 2. Resources tables monitor the usage of tokens, requests, and storage to ensure that the user has enough of them to run a particular model. The Parameters tables define settings for AI models, including temperature and max tokens for Large Language Models. Relationships are established between these tables in a single PocketBase instance for data management, with users linked to their respective resources and files. Corporate or individual specifics being tracked separately. The PocketBase schemas for chats, models, and users align with this structure to store additional details like chat history, model information, and extended user data, including subscription details and usage limits. The key moment for this process is that it happens all at once before sending a request to the model. This allows the application to forward it to the best possible scenario, like in Fig. 4.

### Functions

- **Multiple AI Models:** Artifimo offers a diverse selection of AI models that users can interact with [10]. Each model has its own unique capabilities, knowledge base and personality, different dataset and direction, providing users with a model for every task.
- **Real-Time Chat:** users can participate in real-time conversations with AI models. The chat interface is designed to be responsive and user-friendly on any device.
- **Model Selection:** Artifimo provides a convenient model selection feature, allowing users to easily switch between different AI models for each chat. Users can explore and compare the responses and capabilities of different models to find the one that best suits their needs.
- **Settings and Control:** the platform offers

settings for tokens, temperature, prompts and creativity level.

- **Message History:** Artifimo keeps a record of the history of conversations between the user and each AI model. Users can easily go back to previous messages and continue the conversation up to the point where they left off.
- **User Authentication:** Artifimo includes user authentication and authorization features to provide secure access to the platform. Users can create accounts, log in and manage their accounts.
- **File attachment:** there is an option to attach a file to a chat with the goal that the model can work with the user's data. This allows the user to answer questions about the document, rewrite it, etc. Also known as RAG (Retrieval-Augmented Generation).
- **Internet access:** there is an option via SerpApi [11] for models to access the Internet and view web pages in real time. This overcomes the hurdle of the model having limited knowledge of new events.
- **Autonomous UI Component Creation:** by enabling the "Interactive Mode" option, the AI can decide when it is appropriate to create and display a UI component. This is useful for displaying information such as weather, flight status, merchandise information, etc.
- **API Integration:** The platform integrates with an external API to retrieve information about available AI models and their configurations.

### CONCLUSIONS

Artifimo is a powerful and easy-to-use platform that allows users to engage in interactive conversations with a variety of AI models, according to the task at hand. With an intuitive interface, customizable settings, and the ability to chat in real-time, upload and train your own models, Artifimo is the future of AI chat. Artifimo brings all these tasks together in a single, easy-to-use application. The user interface (UI) of the Artifimo web application is demonstrated within the publication.

## REFERENCES

1. IBM. <https://newsroom.ibm.com/2022-05-19-Global-Data-from-IBM-Shows-Steady-AI-Adoption-as-Organizations-Look-to-Address-Skills-Shortages,-Automate-Processes-and-Encourage-Sustainable-Operations>, Available 19 May 2022.
2. Radix UI. <https://www.radix-ui.com/>
3. Zustand. <https://github.com/pmndrs/zustand>, Available June 2019.
4. Pocketbase. <https://pocketbase.io/>
5. Vercel - Hosting platform. <https://vercel.com>
6. WebLLM - In-browser LLM inference engine that brings language model inference directly onto web browsers with hardware acceleration. <https://github.com/mlc-ai/web-llm>, Available April 2023.
7. PyTorch Python Library. <https://pytorch.org/>
8. HuggingFace - A company providing tools and libraries for natural language processing. <https://huggingface.co>
9. AdamW Optimizer - A variant of the Adam optimizer in Machine Learning that separates weight decay from the gradient update. <https://huggingface.co/docs/bitsandbytes/main/en/reference/optim/adamw>, Available 10 April 2024.
10. Artifimo Model Catalog. <https://artifimo.com/models>, Available 17 April 2024.
11. SerpApi - API for Internet access. <https://serpapi.com>
12. Google OAuth - Used to authenticate users through their Google accounts. <https://developers.google.com/identity/protocols/oauth2>
13. Next.js - For building React applications. <https://nextjs.org>