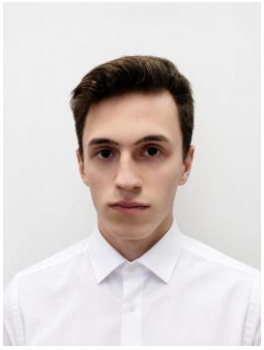
DOI https://doi.org/10.46845/2541-8254-2026-4(56)-71-71
УДК 004.8

АРХИТЕКТУРА ЭКОСИСТЕМЫ ЛОКАЛЬНОГО ВЗАИМОДЕЙСТВИЯ ПОЛЬЗОВАТЕЛЯ С БОЛЬШИМИ ЯЗЫКОВЫМИ МОДЕЛЯМИ НА ПРИМЕРЕ ПЛАТФОРМЫ BRAINE



В. И. Свечников, студент группы 23ВТ2
E-mail: electrowaib@gmail.com
ФГБОУ ВО «Калининградский государственный технический университет»

Е. Ю. Заболотнова, канд. пед. наук
E-mail: elena.zabolotnova@klgtu.ru
ФГБОУ ВО «Калининградский государственный технический университет»

В статье рассмотрена архитектура экосистемы BRAINE, предназначенной для персонализированного взаимодействия пользователя с большими языковыми моделями (LLM). Предложен гибридный подход, объединяющий веб-платформу с голосовым чатом на базе Gemini API и десктопное приложение для локального инференса моделей формата GGUF через llama.cpp. Описаны ключевые компоненты системы: модуль Live2D-визуализации с lip-синхронизацией, режим «окно в окне» с захватом экрана, система кастомизации персонажей и фонов. Приведены результаты сравнительного анализа с существующими аналогами и экспериментальные замеры производительности моделей семейства Qwen 2.5. Показано, что локальный инференс обеспечивает скорость до 120 токенов/с на GPU для модели 3B при отсутствии зависимости от облачных сервисов и подписок. Обсуждены перспективы применения системы в качестве альтернативы традиционным горячим линиям и инструментам повышения цифровой доступности.

Ключевые слова: *большие языковые модели, локальный инференс, llama.cpp, GGUF, экосистема взаимодействия человек-компьютер, Live2D, голосовой интерфейс, цифровая доступность.*

ВВЕДЕНИЕ

Стремительное развитие генеративного искусственного интеллекта (ИИ) привело к широкому распространению диалоговых агентов на базе больших языковых моделей (LLM). При этом современные LLM обладают как значительным прикладным потенциалом, так и рядом ограничений, связанных с надежностью, интерпретируемостью и рисками некорректного использования результатов генерации [1]. Однако подавляющее большинство коммерческих решений (ChatGPT, Character.AI, Replika) требует постоянного подключения к облачным сервисам, что порождает ряд проблем: утечка персональных данных, зависимость от доступности зарубежной инфраструктуры, необходимость платных подписок и цифровое неравенство для жителей регионов с нестабильным интернетом [2].

Альтернативой являются локальные решения (LM Studio, Oobabooga Text Generation WebUI), позволяющие запускать открытые модели на собственном оборудовании и снижать зависимость пользователя от удаленной инфраструктуры [3, 4]. Однако такие инструменты, как правило, ориентированы на технически подготовленных пользователей, лишены визуального персонажа-спутника и не интегрируются с повседневным рабочим процессом пользователя, включая игры, браузер, IDE и другие приложения.

Целью настоящего исследования является разработка и обоснование архитектуры экосистемы BRAINE, сочетающей доступность облачных сервисов с приватностью и незави-

симостью локального инференса, а также визуальное персонажное взаимодействие с интеграцией в повседневную деятельность пользователя на основе современных инструментов запуска LLM, голосового интерфейса и Live2D-визуализации [5–7].

ОБЪЕКТ ИССЛЕДОВАНИЯ

Объектом исследования является процесс взаимодействия конечного пользователя с большими языковыми моделями в контексте персонализированного цифрового сопровождения.

Предметом исследования выступает архитектура программной экосистемы BRAINE, включающей: веб-платформу BRAINE WEB с голосовым чатом и Live2D-визуализацией; десктопное приложение BRAINE DESKTOP для локального инференса; единую систему кастомизации, синхронизации и адаптивного интерфейса.

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Целью работы являлась разработка архитектуры экосистемы BRAINE для локального и гибридного использования больших языковых моделей, обеспечивающей персонализированное взаимодействие пользователя с ИИ-ассистентом через текстовый, голосовой и визуальный интерфейсы. Достижение указанной цели предполагало не только проектирование отдельных программных модулей, но и их объединение в единую среду, в которой пользователь может выбирать между облачным сценарием работы, обеспечивающим быстрый доступ к мультимодальным возможностям, и локальным сценарием инференса, ориентированным на приватность, автономность и отсутствие зависимости от подписочных сервисов.

К задачам исследования относятся анализ существующих решений для взаимодействия с большими языковыми моделями и выявление их ограничений с точки зрения приватности, доступности, персонализации и интеграции в пользовательский рабочий процесс; проектирование модульной архитектуры, объединяющей веб-компонент BRAINE WEB и десктоп-компонент BRAINE DESKTOP; определение состава функциональных модулей, отвечающих за голосовой ввод, захват экрана, визуальную обратную связь через Live2D-аватар и настройку пользовательского окружения.

Кроме того, к задачам исследования относятся реализация и описание механизмов локального запуска моделей формата GGUF через llama.cpp; проведение экспериментального тестирования производительности моделей семейства Qwen 2.5 в условиях пользовательского аппаратного стенда; сопоставление разработанной экосистемы с существующими аналогами; оценка перспектив применения BRAINE для повышения цифровой доступности, персонального сопровождения пользователя и возможной замены части функций традиционных горячих линий.

МЕТОДЫ ИССЛЕДОВАНИЯ

В работе применены методы системного анализа, объектно-ориентированного проектирования и экспериментального тестирования программного обеспечения.

Анализ предметной области проводился на основе обзора научной литературы и технической документации, включая материалы по llama.cpp, Gemini API и Live2D Cubism SDK [5–7]. Архитектура системы проектировалась с использованием декомпозиции на независимые модули с четко определенными интерфейсами, что позволило разделить функции локального инференса, голосового взаимодействия, визуализации и пользовательской кастомизации.

Локальный инференс реализован на базе библиотеки llama.cpp с поддержкой формата GGUF и GPU-ускорения через CUDA [5]. Веб-компонент использует Gemini Live API для многомодального голосового диалога [6]. Визуализация персонажей построена на Live2D Cubism Core с кастомным рендерером на WebGL/Canvas [7].

Тестирование производительности проводилось на стенде: Intel Core i5-12400, 32 ГБ DDR4, NVIDIA GTX 1660 6 ГБ, Windows 11. Замеры скорости генерации выполнялись

встроенными средствами llama.cpp с параметром полного оффлоудинга на GPU. Дополнительно проверялись функциональные сценарии BRAINE Desktop: запуск GGUF-модели, потоковый вывод ответа, прикрепление файлов и отображение Markdown в пользовательском окне программы.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Архитектура экосистемы

Разработанная экосистема BRAINE состоит из двух взаимосвязанных компонентов (рисунок 1).

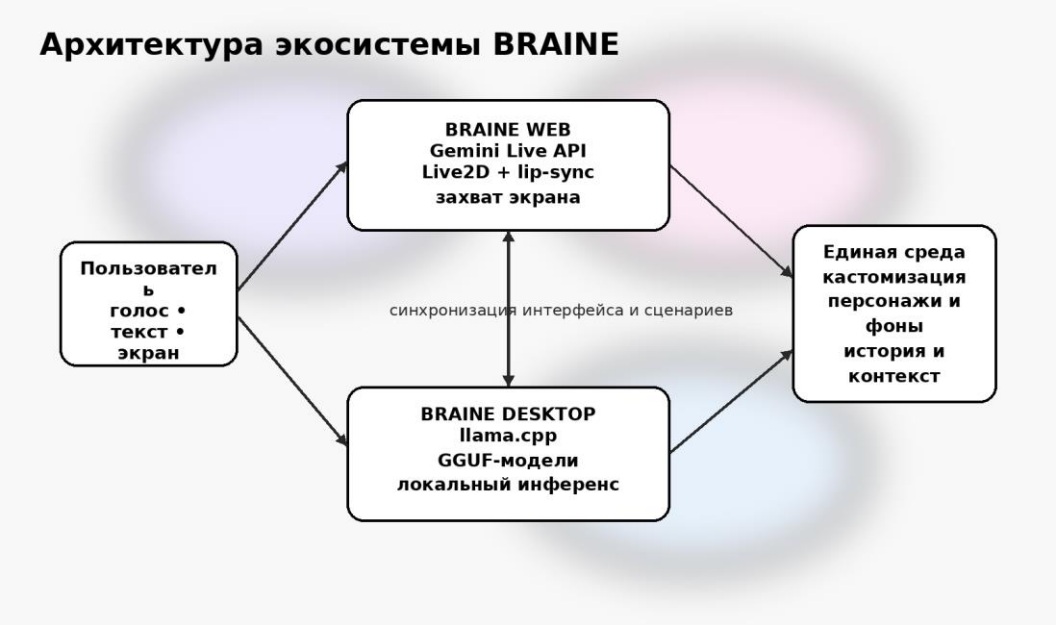


Рисунок 1 – Общая архитектура экосистемы BRAINE

BRAINE WEB представляет собой полностью адаптивную веб-платформу (360 px – 4K), обеспечивающую голосовой чат с AI-персонажем через Gemini Live. Персонаж отображается в формате Live2D с автоматической синхронизацией движения губ (lip-sync) с потоковым аудиоответом. Пользователю доступны 6 встроенных персонажей и возможность загрузки собственных моделей с ПК. Реализованы режимы ввода: микрофон, захват экрана, текст, стоп-слово для прерывания диалога. Схема пользовательского окна BRAINE WEB с областью диалога, панелью персонажей и Live2D-аватаром представлена на рисунке 2.



Рисунок 2 – Пользовательское окно BRAINE WEB с областью диалога и встроенными персонажами

BRAINE DESKTOP – нативное приложение на .NET 8 для Windows 10/11, macOS и Linux. Минимальное разрешение окна 520×420 px обеспечивает работу на слабых машинах. Приложение запускает GGUF-модели локально через llama.cpp без облака и подписок. Реализован стриминг ответов с анимацией печатной машинки, полный рендеринг Markdown (код с подсветкой, жирный текст, блоки копирования). Поддерживается прикрепление изображений, PDF, кода и текстовых файлов с автоматическим извлечением текста до 24 000 символов контекста. Реализация этих функций в пользовательском окне BRAINE Desktop показана на рисунке 3.



Рисунок 3 – Пользовательское окно BRAINE Desktop со стримингом ответа и прикреплением файлов

Сравнительный анализ

Проведен сравнительный анализ с аналогами трех типов: коммерческими персонажными сервисами (Character.AI и Replika), локальными инструментами запуска языковых моделей (LM Studio и Oobabooga Text Generation WebUI), а также гибридным персонажным интерфейсом SillyTavern, который может работать с локальным self-hosted LLM-backend [2–4, 8]. Такой выбор обусловлен тем, что перечисленные решения покрывают основные оси сравнения: персонажное взаимодействие, локальный запуск моделей, голосовой интерфейс, интеграцию с пользовательским рабочим процессом и независимость от подписочных облачных сервисов.

Существование гибридных решений не позволяет считать корректным абсолютное утверждение об отсутствии аналогов. Поэтому в таблице 1 сравнение проводится не с «аналогами» в целом, а с репрезентативными решениями, частично пересекающимися с BRAINE по отдельным функциям.

Анализ показал, что среди рассмотренных решений BRAINE не является единственным проектом, в котором присутствуют отдельные признаки персонажного интерфейса и локального запуска. Однако отличием BRAINE является объединение в одной экосистеме локального инференса, Live2D-персонажа с lip-sync, голосового взаимодействия, режима «окно в окне», захвата экрана и сценариев локальной/корпоративной интеграции. Следовательно, корректнее говорить не об отсутствии аналогов как таковых, а о недостатке решений, совмещающих все перечисленные функции в едином пользовательском сценарии.

Таблица 1 – Сравнительные характеристики экосистемы BRAINE и репрезентативных аналогов

Критерий	BRAINE	Character.AI	Replika	LM Studio	Oobabooga	SillyTavern
Локальный инференс	+	–	–	+	+	±
Персонажное взаимодействие	+	+	+	–	–	+
Голосовой чат / lip-sync	+	±	±	–	–	±
Live2D или визуальный аватар	+	–	±	–	–	±
Режим «окно в окне»	+	–	–	–	–	–
Захват экрана	+	–	–	–	–	–
Открытая загрузка моделей	+	–	–	+	+	±
Без подписок при локальном сценарии	+	–	–	+	+	+
Локальная/корпоративная интеграция	+	–	–	±	+	±

Примечание. «+» – реализовано; «–» – не реализовано; «±» – реализуется частично или требует внешнего компонента.

Режим «окно в окне»

Ключевой особенностью, выделяющей BRAINE на фоне рассмотренных решений, является режим «окно в окне» (рисунок 4). Персонаж выносится в отдельное плавающее окно, остающееся поверх вкладок, игр и приложений. Окно перетаскивается в любую точку экрана, изменяет размер от 480 до 720 px, не мешает основному рабочему процессу и не исчезает при переключении между приложениями. Под «непрерывным присутствием AI-ассистента» в данной работе понимается постоянная визуальная и интерфейсная доступность ассистента для пользователя, а не непрерывный скрытый анализ экрана или звука. Контекстный анализ выполняется только при явном включении пользователем захвата экрана, микрофона или текстового ввода.



Рисунок 4 – Примеры работы режима «окно в окне» в пользовательских сценариях

Результаты тестирования локального инференса и интерфейса BRAINE Desktop

Проведено тестирование производительности моделей семейства Qwen 2.5 в формате GGUF (таблица 2, рисунок 5). Линейка Qwen 2.5 выбрана как один из актуальных классов открытых языковых моделей для локального запуска, имеющий несколько размерностей (3B, 7B, 14B) и позволяющий оценить компромисс между скоростью генерации, потреблением памяти и предполагаемым качеством ответа на одном аппаратном стенде [9]. В рамках эксперимента сравнивались только модели одной линейки, поэтому выводы не распространяются автоматически на другие семейства LLM.

Для разграничения тестирования модели и тестирования интерфейса отдельно проверялись пользовательские сценарии BRAINE Desktop: загрузка GGUF-файла, вывод ответа в потоковом режиме, обработка прикрепленных файлов и корректный рендеринг Markdown. Прямое сопоставление скорости той же модели в LM Studio или Oobabooga требует дополнительного эксперимента с идентичными параметрами контекста, квантизации, числа потоков и GPU-offload; в текущей версии работы это ограничение фиксируется явно.

Таблица 2 – Производительность локального инференса

Модель	CPU, ток/с	GPU (CUDA), ток/с	Память, ГБ
Qwen 2.5 3B	45	120	2.1
Qwen 2.5 7B	28	85	4.8
Qwen 2.5 14B	12	42	9.2

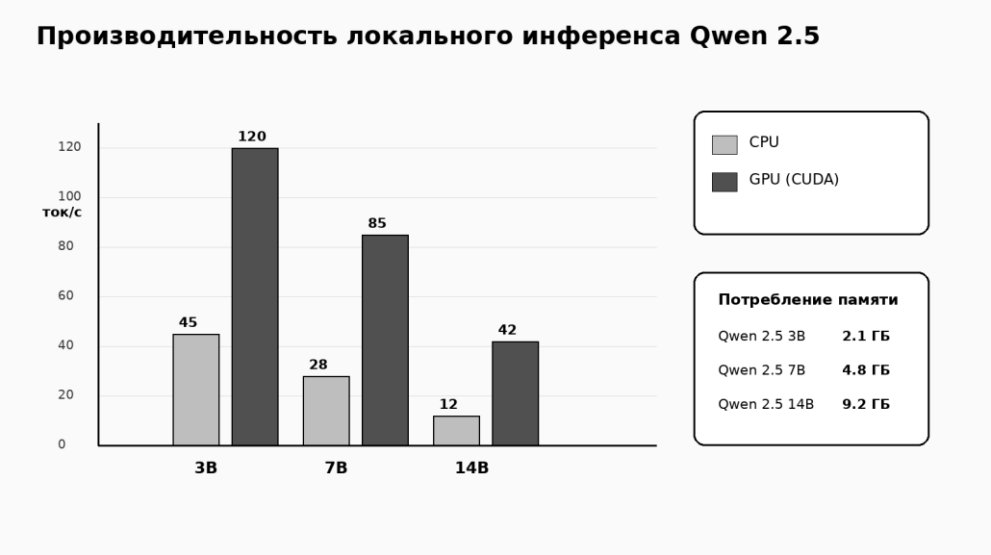


Рисунок 5 – Производительность локального инференса моделей Qwen 2.5 в приложении BRAINE Desktop

Среди протестированных моделей семейства Qwen 2.5 модель 3B демонстрирует наибольшую скорость (120 ток/с на GPU) при минимальном потреблении памяти (2.1 ГБ), что делает её наиболее сбалансированным вариантом для слабых машин в пределах выбранной линейки. Модель 14B обеспечивает потенциально более высокое качество ответов при сохранении приемлемой скорости 42 ток/с на GPU среднего класса, но требует существенно большего объема памяти. Таким образом, вывод об оптимальном выборе Qwen 2.5 3B ограничен только экспериментом с линейкой Qwen 2.5 и не исключает, что другие версии Qwen или модели иных семейств могут оказаться предпочтительнее при других аппаратных условиях и требованиях к качеству ответа.

Таблица 3 – Проверка функциональных сценариев BRAINE Desktop

Сценарий	Проверяемый признак	Результат
Запуск GGUF-модели	Модель загружается локально через llama.cpp без облачного API	Выполнено
Стриминг ответа	Текст появляется постепенно в пользовательском окне, имитируя печать	Выполнено
Прикрепление файлов	Изображения, PDF, код и текстовые файлы добавляются к запросу с извлечением текста	Выполнено
Рендеринг Markdown	Заголовки, кодовые блоки, жирный текст и копирование блоков отображаются корректно	Выполнено

Кастомизация и доступность

Реализована система кастомизации: позиция и размер персонажа настраиваются слайдерами X, Y и Scale в реальном времени; персонаж следит за курсором и реагирует на взаимодействие. Поддерживается загрузка произвольного фона с ПК или по ссылке для создания игровой сцены.

Адаптивность интерфейса обеспечивает работу на устройствах от 360 px до 4K. Голосовое управление, простой интерфейс и отсутствие необходимости установки (веб-версия) делают систему доступной для пожилых людей, лиц с ограниченными возможностями и жителей регионов.

ЗАКЛЮЧЕНИЕ

В ходе исследования разработана архитектура экосистемы BRAINE, объединяющей веб-платформу с облачным голосовым чатом и десктопное приложение для полностью локального инференса LLM. Полученные результаты подтверждают возможность объединения облачных API, локальных GGUF-моделей и визуального персонажного интерфейса в единую пользовательскую среду [5–7]. Ключевые результаты:

1. Предложен гибридный подход, позволяющий пользователю выбирать между облачным удобством (Gemini) и полной приватностью (локальные GGUF-модели через llama.cpp).

2. Реализован режим «окно в окне» с захватом экрана по инициативе пользователя, обеспечивающий постоянную интерфейсную доступность ассистента и контекстную работу без непрерывного скрытого анализа экрана или звука.

3. Показана практическая применимость локального инференса на выбранном стенде: среди протестированных моделей Qwen 2.5 модели 3B/7B достигают 120/85 ток/с на GPU среднего класса.

4. Показан потенциал системы для повышения цифровой доступности и в качестве государственной альтернативы традиционным горячим линиям — достаточно заменить Gemini API на ГигаЧат, YandexGPT или любую локальную модель, чтобы данные не покидали страну.

Перспективы развития включают интеграцию с российскими LLM (GigaChat, YandexGPT), расширение библиотеки персонажей, добавление мультиязычного распознавания речи и развитие механизмов долгосрочной памяти диалога. К направлениям дальнейшего исследования также относится сравнительное тестирование BRAINE Desktop с другими интерфейсами запуска моделей (LM Studio, Oobabooga) и с моделями иных семейств при одинаковых параметрах контекста, квантизации и GPU-offload.

СПИСОК ЛИТЕРАТУРЫ

1. Bender, E. M. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? / E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. – 2021. – P. 610–623. – DOI: 10.1145/3442188.3445922. – Режим доступа: https://dl.acm.org/doi/10.1145/3442188.3445922 (дата обращения: 06.05.2026).

2. Character.AI. Platform overview [Электронный ресурс]. – Режим доступа: https://character.ai (дата обращения: 03.05.2026).

3. LM Studio. Local LLM inference [Электронный ресурс]. – Режим доступа: https://lmstudio.ai (дата обращения: 03.05.2026).

4. Oobabooga. Text Generation WebUI / textgen [Электронный ресурс]. – Режим доступа: https://github.com/oobabooga/textgen (дата обращения: 06.05.2026).

5. ggml-org. llama.cpp: LLM inference in C/C++ [Электронный ресурс]. – Режим доступа: https://github.com/ggml-org/llama.cpp (дата обращения: 06.05.2026).

6. Google DeepMind. Gemini API documentation [Электронный ресурс]. – Режим доступа: https://ai.google.dev/gemini-api/docs (дата обращения: 03.05.2026).

7. Live2D Inc. Cubism Core. SDK Manual [Электронный ресурс]. – Режим доступа: https://docs.live2d.com/en/cubism-sdk-manual/cubism-core/ (дата обращения: 06.05.2026).

8. SillyTavern. What is SillyTavern? [Электронный ресурс]. – Режим доступа: https://docs.sillytavern.app/ (дата обращения: 13.06.2026).

9. Qwen series [Электронный ресурс] / Alibaba Cloud. – Режим доступа: https://qwenlm.github.io (дата обращения: 03.05.2026).

ARCHITECTURE OF THE ECOSYSTEM FOR LOCAL USER INTERACTION WITH LARGE LANGUAGE MODELS: THE BRAINE PLATFORM CASE STUDY

V. I. Svechnikov, student of group 23VT2
E-mail: electrowaib@gmail.com
Kaliningrad State Technical University

E. Y. Zabolotnova, Candidate of Pedagogical Sciences, Associate Professor
E-mail: elena.zabolotnova@klgtu.ru
Kaliningrad State Technical University

The paper examines the architecture of the BRAINE ecosystem designed for personalized user interaction with large language models (LLM). A hybrid approach has been proposed, combining a web platform with voice chat based on Gemini API and a desktop application for local inference of GGUF-format models via llama.cpp. Key system components have been described: Live2D visualization module with lip synchronization, «window-in-window» mode with screen capture, character and background customization system. The results of comparative analysis with existing analogues and experimental measurements of Qwen 2.5 family models performance have been presented. It has been shown that local inference provides speed up to 120 tokens/sec on GPU for the 3B model without dependence on cloud services and subscriptions. The prospects of using the system as an alternative to traditional hotlines and tools for improving digital accessibility have been discussed.

Keywords: *large language models, local inference, llama.cpp, GGUF, human-computer interaction ecosystem, Live2D, voice interface, digital accessibility.*