ИССЛЕДОВАНИЕ ПРОИЗВОДИТЕЛЬНОСТИ ГЕНЕРАТИВНОГО ИИ НА ВСТРАИВАЕМЫХ ARM-ПЛАТФОРМАХ ПОД УПРАВЛЕНИЕМ ОС «АВРОРА»

Гуненков М.Ю., Цейдлер А.А., Фральцов М.А.
Омский государственный технический университет, г. Омск, пр-кт Мира, д. 11
myugunenkov@omgtu.ru
anzej2003@gmail.com
matvey.fraltsov@mail.ru

Актуальность. Развитие отечественных мобильных платформ и технологий генеративного искусственного интеллекта приводит к задаче оценки возможности локального развёртывания больших языковых моделей (LLM) непосредственно на устройствах без необходимости обращения к внешним облачным сервисам. Известно, что такой подход позволяет обеспечивать конфиденциальность данных, независимость от сетевого подключения, а также снижает зависимость от зарубежной инфраструктуры. Вместе с тем аппаратные ограничения мобильных устройств, такие как ограниченный объём оперативной памяти, относительно низкая вычислительная мощность ARM-процессоров и тепловые ограничения, требуют поиска компромисса между пропускной способностью генерации и потреблением ресурсов. Систематическое измерение и сравнение производительности моделей на устройствах под управлением ОС Аврора является актуальной научно-практической задачей.

Цели и задачи. Целью работы является проведение комплексного тестирования производительности инференса языковых моделей различной архитектуры и уровня квантования на устройствах под управлением ОС Аврора для определения конфигураций, обеспечивающих приемлемую скорость отклика при русскоязычном пользовательском сценарии. Задачи включают разработку инструментальной цепочки для воспроизводимого автоматизированного сбора данных, инференс моделей на целевых устройствах и последующий статистический анализ влияния схемы квантования, числа вычислительных потоков и аппаратной платформы на производительность моделей.

Методология. Для решения поставленных задач разработан специализированный инструментальный стек. Основанная на библиотеке llama.cpp утилита «Inferencer» выполняет локальный инференс GGUF-моделей и формирует структурированный файл в формате JSON с детализированными метриками: время загрузки модели и генерации токенов, потребление оперативной памяти, загрузка и температура CPU. Написанный на языке Go модуль-оркестратор «Bencher» автоматически разворачивает матрицу сценариев по моделям, схемам квантования, числу потоков, языкам и числу повторов, агрегируя результаты в CSV-файл. Совокупность инструментов создаёт воспроизводимую методику тестирования производительности инференса LLM на целевых устройствах под управлением ОС Аврора. Исследование проводилось на устройствах TrustPhone T1 (MediaTek Helio P70, ОС Аврора 5.1) и Kvadra T (архитектура Cortex-A55, ОС Аврора 5.2). Тестировались пять языковых моделей (Qwen2.5 0.5B, Vikhr-Qwen2.5 0.5B, Llama3.2 1B, Gemma4 2B и GigaChat3-pruned 5B) в восьми режимах квантования (IQ4_XS, Q8_0, Q5_K_M, Q4_K_M, Q2_K, Q2_K_XL, Q8_K_XL, и I1_Q6_K). Все запуски производились с активированной NEON-оптимизацией при числе потоков 2, 4, 6 и 8. Применялись следующие сценарии: Minimum (генерация 50 токенов), Baseline UX (128 токенов), Baseline Throughput (230 токенов) и Long Context (128 токенов при расширенном контексте). Генерация осуществлялась на русском и английском языках. Итоговый набор данных содержит 2626 измерений, каждое из которых включает 40 исследуемых параметров. Для статистической проверки применялись доверительные интервалы и перестановочные тесты.

Результаты. Модели Qwen2.5 и Vikhr-Qwen2.5 лидируют по скорости генерации на обоих устройствах. На Kvadra T средний TPS составил: Qwen2.5 – 18.68 ток/с ($\sigma = 4.90$), Vikhr-Qwen2.5 – 15.63 ток/с ($\sigma = 3.63$), Llama3.2 – 8.21 ток/с ($\sigma = 2.08$). На TrustPhone T1: Vikhr-Qwen2.5 – 10.83 ток/с ($\sigma = 3.41$), Qwen2.5 – 10.63 ток/с ($\sigma = 2.51$), Llama3.2 – 4.78 ток/с ($\sigma = 1.17$). Модели Gemma4 и GigaChat3-pruned показали значительно более низкие значения

TPS. Статистически подтверждена практическая граница пиковой нагрузки на оперативную память. Показано, что конфигурации с потреблением до 1000 МБ превосходят конфигурации с большим потреблением как по скорости генерации, так и по времени получения первого токена. В этот диапазон попадают все схемы квантования для моделей Qwen2.5 и Vikhr-Qwen2.5, а также схемы IQ4_XS, Q4_K_M и Q5_K_M для Llama3.2. При этом масштабирование по числу потоков носит устройство-зависимый характер. Для TrustPhone T1 прирост производительности сохраняется вплоть до 8 потоков, тогда как для Kvadra T переход от 6 к 8 потокам у части моделей не даёт выигрыша в производительности. Установлено, что для полноценной оценки пользовательского опыта необходимо дополнять TPS показателем WPS (количество слов в секунду). Модель Vikhr-Qwen2.5 обеспечивает WPS на русском языке на уровне английского и выше благодаря эффективной токенизации, в то время, как у Qwen2.5 и Llama3.2 значения WPS на русском заметно ниже при сопоставимом TPS. Обнаружено, что между Qwen2.5 и Vikhr-Qwen2.5 отсутствует универсальное ранжирование. На Kvadra T средний TPS у Qwen2.5 статистически выше, тогда как на TrustPhone T1 разница между моделями статистически неубедительна.

Заключение. Проведённое исследование показывает, что локальный инференс LLM на устройствах с ОС Аврора обеспечивает целевые показатели скорости отклика при пиковом потреблении оперативной памяти до 1000 МБ. Модели семейств Qwen2.5 и Vikhr-Qwen2.5 с квантованием Q4_K_M или IQ4_XS обеспечивают скорость генерации выше 8–10 ток/с на современных ARM-платформах, что является достаточным показателем для базовых пользовательских сценариев. В то же время при акценте на русскоязычный контент предпочтительно использование модели Vikhr-Qwen2.5. Разработанный инструментальный стек обеспечивает воспроизводимость экспериментов и может применяться для тестирования новых моделей и устройств под управлением ОС Аврора.