УДК: 004.051, 004.057, 004.08.

Игорь Викторович Потеряхин, студент кафедры Б6,
Анна Дмитриевна Тифрани, студентка кафедры Б6,
Егор Юрьевич Круглов, студент кафедры Б6.
Научный руководитель: **Анастасия Анатольевна Волкова**., кандидат экономических наук, доцент кафедры Б6 БГТУ «Военмех» им. Д.Ф.Устинова.

Федеральное государственное бюджетное образовательное учреждение высшего образования «Балтийский государственный технический университет „ВОЕНМЕХ“ им. Д.Ф. Устинова», г. Санкт-Петербург, Российская Федерация.

ИЗУЧЕНИЕ СКОРОСТИ ИНФЕРЕНСА ЯЗЫКОВЫХ МОДЕЛЕЙ

Аннотация: Статья посвящена исследованию локального инференса больших языковых моделей на бытовом и серверном оборудовании с использованием LM Studio. Протестированы пять конфигураций: от процессора Xeon без AVX2 до i7-11700k с RTX 3080Ti и стенда с 72 ГБ VRAM на Tesla P40. Оценены скорости генерации моделей от 8 В до 120 В параметров, влияние размера VRAM, архитектуры моделей (Dense vs MoE, GQA), квантования и пропускной способности памяти. Сделан вывод, что для эффективного домашнего инференса критически важен объём видеопамяти и скорость обмена с ней, а частота видеопроцессора или объём ОЗУ - второстепенны.

Ключевые слова: локальный инференс, формула скорости инференса, большие языковые модели (LLM), архитектура моделей, видеокарты, скорость генерации (токен/с), аппаратные ограничения инференса.

1. Введение: Почему локальность — это новая норма

Эпоха слепого доверия облачным API постепенно уступает место культуре локального запуска LLM. Причины очевидны: приватность данных, отсутствие цензуры и, в долгосрочной перспективе, экономия. Однако запуск современной

языковой модели (LLM) на бытовом или списанном серверном железе — это всегда битва за токены в секунду (t/s). В данной статье рассматривается работа популярных моделей (от Llama 3 до Nemotron) через призму LM Studio. Будет протестировано пять уникальных конфигураций — от «музейного» Xeon до современного домашнего i7 с RTX 3080Ti — чтобы понять, где на самом деле находится «бутылочное горлышко» современного инференса.

Для достижения поставленной цели будут выполнены следующие задачи:

1) Отобрать большие языковые модели и исследовать их характеристики;

2) Запустить модели, после чего оценить результаты их работы;

3) Провести формулирование недостатков и преимуществ каждой модели и сравнение между ними;

4) Определить, имеется ли закономерность в скорости генерации от особенностей языковой модели;

5) Кратко проанализировать более тяжёлые (или объёмные) языковые модели;

2. Лаборатория: Участники эксперимента

Для анализа тестовые стенды были разделены на две категории. Первые четыре — это типичные представители «домашнего» или «серверного» инференса с ограниченным объёмом VRAM. Пятый стенд — экспериментальный полигон для тяжёлых моделей.

На изображении 1 отражена Таблица 1: Рассматриваемые стенды и описание их характеристик.

Группа А: Основной состав

№	Стенд	Процессор	Память	Видеокарта (VRAM)	Особенности
1	CPU King	E5-2696v3	128G Quad DDR4-2133	GTX1080 8G	Сток, упор на CPU-инференс
2	Old School	E5-2689v4	32G Dual DDR4-2400	2x1080Ti 22G	Pascal-связка, большой объём VRAM
3	Modern	i7-11700k	128G Quad DDR4-2933	3080Ti 12G	Высокая частота ядер, архитектура Ampere
4	Museum	2xE5-2697v2	384G 2xQuad DDR3-1066	—	Нет AVX2, огромный объём медленной RAM, только CPU

Группа Б: Спецзадание (VRAM-Монстр)

№	Стенд	Процессор	Память (RAM)	Видеокарта (VRAM)	Кейс исследования
5	The Monster	E5-2696v3	128G DDR4 Quad	3x Tesla P40 72G	72GB VRAM

Изображение 1

Аналитический акцент:

Разделение на эти группы позволяет сначала рассмотреть «бытовые» проблемы (когда модель не вмещается в видеокарту и тормозит), а затем, на примере Стенда №5, показать, чего можно достичь, когда имеются 72 ГБ видеопамяти и модель гарантированно сидит в VRAM.

Экономика инференса: Почему X99 — фаворит

Выбор платформ Haswell-E и чипсета X99 в данном тестировании не случаен. В мире локального ИИ цена одного токена в секунду напрямую зависит от стоимости подсистемы памяти. Авторы намеренно игнорируют современные решения на DDR5: их стоимость сопоставима с бюджетом всей остальной сборки, при этом реальный прирост в задачах инференса не оправдывает затрат(нет роста скорости пропорционально росту цены). В то же время классический Dual Channel DDR4 на потребительских платформах (как в нашем стенде #3) проигрывает серверным решениям по соотношению «цена/пропускная способность». Основной тезис данной статьи: DDR4-2133 в четырёхканальном режиме (Quad Channel) на платформе X99 — это «золотая середина» 2026 года. Она обеспечивает достаточную пропускную способность для работы с весами моделей, которые не уместились в видеопамять, оставаясь при этом самым бюджетным способом получить 128 ГБ оперативной памяти и

выше. Особенно при современных ценах на память и относительной доступности DDR4 REG ECC памяти.

В данном исследовании не проводится анализ качества ответов сетей, так как это выходит за рамки статьи. Но разумеется, что модель с большим числом параметров даёт лучше ответы, и по логике, и по широте знаний. По мнению авторов, реальность такова, что качественный ответ на вопрос, особенно если это решение какой-то технической задачи с использованием инструментов - это модель от 30B (миллиардов) параметров.

3. Проклятие медленной памяти: почему 384 ГБ не спасают

На стенде #5 (Museum) обнаруживается главное заблуждение новичков: «Куплю по дешёвке сервер с горой оперативки и запущу 70B модель». Результат в 4-5 t/s на всех, даже небольших, моделях — это приговор медленной DDR3-1066. В объёме нет смысла, если нет скорости. Даже работая в «двойном» четырёхканальном режиме, эта память не способна обеспечить поток данных, необходимых процессору. Инференс LLM — это задача, предельно чувствительная даже не столько к задержкам, сколько к пропускной способности памяти. Пока веса модели «едут» из медленной RAM в процессор по узкой шине десятилетней давности, вычислительные ядра простаивают. Здесь даже добавление современных инструкций AVX2 не спасло бы ситуацию: процессор просто «голодает», ожидая данные. Это наглядный урок: в «домашнем» инференсе скорость памяти важнее её объёма. Да, как-то ещё шевелится Nemotron Nano в кванте Q3, но сам смысл 384Гб памяти потерян.

4. «VRAM-обрыв»: Жёсткая бинарность производительности

Анализ результатов стендов #2 (Old School) и #3 (Modern) выявляет самую драматичную закономерность данного исследования. В мире GPU-инференса не существует «почти влезло».

Рассмотрим Nemotron 30B (Q3_K_L) весом 15.3 ГБ:

Стенд #3 (3080Ti 12GB): Модель не помещается в видеопамять. LM Studio вынуждена выгружать ~4 ГБ в системную RAM. Итог — падение до 16.97 t/s. Скорость инференса здесь ограничена скоростью шины PCIe и самой медленной плашки оперативной памяти.

Стенд #2 (2x 1080Ti 22GB total): Несмотря на то, что архитектура Pascal старше на два поколения, связка из двух карт даёт 22 ГБ VRAM. Модель целиком умещается в видеопамять. Итог — 32.59 t/s.

Можно увидеть почти двухкратное превосходство «старичков» над современной картой. Вывод суров: производительность в инференсе падает пропорционально объёму модели, не поместившемуся в VRAM. Как только часть весов покидает пределы видеокарты, мощь тензорных ядер Ampere становится бесполезной, так как они простаивают в ожидании данных из системной памяти.

5. Архитектурный поединок: Вес параметров против «умной» активации

На стенде #3 (3080Ti, 12GB) был зафиксирован аномальный разрыв: Gemma 3 (27B) выдала катастрофические 2.31 t/s, тогда как Nemotron 30B показал 16.97 t/s. Разгадка кроется не только в объёме модели, но и в типе её архитектуры: Gemma 3 (27B) это Dense-архитектура - для генерации каждого токена видеокарта вынуждена «прогонять» через свои вычислительные блоки все 27 миллиардов параметров. Это создаёт колоссальную нагрузку на шину памяти. При объёме VRAM в 12 ГБ модель физически не помещается в видеопамять, и постоянная пересылка тяжёлых пластов данных из системной RAM по шине PCIe превращает инференс в слайд-шоу.

Nemotron 30B (MoE — Mixture of Experts): Несмотря на большее общее число параметров, технология «смеси экспертов» активирует лишь малую часть весов для каждого конкретного запроса. Это значительно снижает объём данных, которые нужно пропустить через ядра процессора в моменте. В сочетании с GQA

(Grouped-Query Attention), которая рачительно расходует память под KV-кеш, Nemotron позволяет системе работать в 7.3 раза быстрее, чем «монолитная» Gemma.

6. Феномен Pascal: Когда объём бьёт технологии

Сравнение #2 (2x 1080Ti) и #3 (3080Ti) на модели Nemotron 30B ставит точку в спорах о «старом железе». 3080Ti (12G): 16.97 t/s. против двух 1080Ti (22G): 32.59 t/s. Современная архитектура Ampere (3080Ti) оказывается почти в два раза медленнее связки десятилетней давности. Причина в жёсткой математике: 30-миллиардная модель (даже в кванте Q3) весит 15.3 ГБ. Она физически не помещается в 12 ГБ современной карты. Как только 3.3 ГБ «вываливаются» в оперативную память, общая скорость системы падает до скорости самой медленной шины в цепочке. 22 ГБ видеопамяти на Pascal-связке позволяют модели целиком находиться «внутри» видеокарт, обеспечивая стабильную и высокую скорость генерации без обращения к системной RAM.

В то же время замер по Phi-4 (15B) на стенде #3 (Modern) показывает феноменальные 62.34 t/s. Это почти в 6 раз быстрее, чем на CPU King.

Здесь мы видим синергию: модель полностью вмещается в 12 ГБ видеопамяти 3080Ti, а современный процессор i7-11700k с высокой частотой на ядро мгновенно справляется с подготовкой промпта и деквантизацией. На старых Xeon даже при наличии GPU скорость часто упирается в то, как быстро CPU может «скармливать» данные видеокарте.

7. Выявление закономерности. Формула скорости генерации

В ходе исследования авторами была эмпирически выведена формула скорости генерации– это «усреднённое» отношение скорости памяти к размеру части модели, лежащей в этой памяти. Скорости не суммируются, они «усредняются» с сильным перекосом в сторону медленного сегмента. Когда модель разделена, видеокарта не может начать вычисления над слоями, которые лежат в RAM, пока они не приедут к ней по шине PCIe.

$$Speed = \frac{1}{\sum \frac{Size_i}{BW_i}}$$

Справедливость данного соотношения может быть проверена читателем непосредственным вычислением.

8. Заключение. Итоговые выводы: Золотые правила «домашнего» инференса

На изображении 2 показана Таблица 2: Результаты измерений скорости генерации при инференсе.

Сборка/Модель	**Qwen3 8B**	**Phi4 15B**	**Llama3 8B**	**Gemma 3 27B**	**Nemotron 30B A3B**
квант	**Q8_0**	**Q4_K_M**	**Q8_0**	**Q4_K_M**	**Q3_K_L**
размер, Гб	**8,11**	**8,43**	**7,95**	**15,30**	**19,32**
2696v3+1080	14,35	10,94	15,2	3,56	15,26
2689v4+2x1080Ti	30,08	24,77	30,08	13,61	32,59
11700k+3080Ti	18,55	62,97	80,30	2,31	16,97
2x2697v2 DDR3	5,5	4,78	5,7	2,44	11,6

Изображение 2

На основе проведённых тестов и анализа пяти различных конфигураций формируется «формула успеха» для сборки локальной станции ИИ в 2026 году:

VRAM — это абсолют. Никакая частота процессора или скорость DDR5 не спасёт вас, если модель не умещается в видеопамять. Эффект «VRAM-обрыва» беспощаден: падение скорости при вытеснении весов в RAM происходит не в процентах, а в разы (как можно было наблюдать на примере 3080Ti и Nemotron).

Архитектура важнее параметров. Количество миллиардов параметров (B) — плохой ориентир. Выбирайте модели с GQA (Grouped-Query Attention) и MoE (Mixture of Experts). Они позволяют получать производительность уровня 30B+ моделей на скоростях, близких к 8B-решениям.

X99 + DDR4 Quad Channel = Best Buy. Это сочетание остаётся непревзойдённым по соотношению «цена/возможности». Четырёхканальный режим DDR4-2133 (а лучше 2400) даёт тот необходимый минимум пропускной

способности, который позволяет «докачивать» веса в GPU без катастрофических задержек.

Избегайте «музейных» решений. Серверы на базе DDR3 (v1/v2) сегодня бесполезны для LM Studio. Медленная шина памяти и отсутствие современных инструкций делают их покупку пустой тратой денег, даже если объём RAM исчисляется сотнями гигабайтов.

Вердикт: Для комфортной работы дома сегодня достаточно связки из E5-2696v3 и пары б/у 1080Ti или Tesla P40. Это «народная» конфигурация, которая позволяет не просто «тыкать» нейросеть палочкой, а полноценно использовать её как рабочий инструмент с длинным контекстом и сложными запросами.

На изображении 3 содержится Таблица 3: Результаты тестирования более тяжёлых моделей

Gemma 3 27B	**Llama 3.3 70B**	**Qwen 3.5 35B A3B**	**NVIDIA Nemotron 120B A12B**	**OpenAI GPT-OSS 120B A5B**
Квант				
Q8_0	**Q6_K**	**Q8_0**	**IQ4_NL**	**MXFP4**
Размер, Гб				
27,53	53,91	35,21	60,06	59,03
Размер контекста при инференсе				
131072	65536	262144	524288	131072
Скорость генерации, токен/с				
8,28	4,05	32.48	14.83	32.58

Изображение 3

В таблице выше протестированы более тяжёлые модели на оборудовании с тремя Tesla P40 с общим объёмом 72Гб VRAM. Это, наверное, является верхним пределом для инференса на домашнем оборудовании в ценовом классе 100+ тыс руб. Модели такого класса запускать на описанном ранее слабом оборудовании совсем не разумно, поэтому результаты не для «соревнования», а, скорее, для ознакомления. Справедливости ради, некоторые большие MoE модели можно «воспроизводить» на слабом железе, к примеру GPT-OSS 120B A5B на 2696v3+1080 выдаёт вполне приличные 8 токен/с, но это, всё таки, немного другая «весовая категория».

Список использованных источников:

1. Лингводидактический потенциал больших языковых моделей (LLM) с открытым исходным кодом на примере Gemma 2. Игорь Маев. 2025. URL: https://www.academia.edu/156505942/Лингводидактический_потенциал_больших_языковых_моделей_LLM_с_открытым_исходным_кодом_на_примере_Gemma_2_The_Language_Teaching_Potential_of_Large_Language_Models_LLMs_with_Open_Source_Code_Using_the_Example_of_Gemma_2_ (дата обращения: 29.04.2026)

2. Обзор по LLM. Блог компании Тензор. 2024. URL: https://habr.com/ru/companies/tensor/articles/790984/ (дата обращения: 29.04.2026)

3. Список графических процессоров Nvidia 2026. URL: https://ru.wikipedia.org/wiki/Список_графических_процессоров_Nvidia (дата обращения: 29.04.2026)

4. DDR3 vs DDR4 vs DDR5: как меняется выбор серверной памяти в 2026 году. Блог Команды пресейла ITELON. 2026. URL: https://itelon.ru/blog/ddr3-vs-ddr4-vs-ddr5/ (дата обращения: 29.04.2026)

5. List of Intel Xeon processors. Wikipedia. 2026. URL: https://en.wikipedia.org/wiki/List_of_Intel_Xeon_processors#Haswell-based (дата обращения: 29.04.2026)

6. LM Studio. 2026. URL: https://lmstudio.ai/ (дата обращения: 29.04.2026)

7. NVIDIA GeForce GTX 1080 Ti Specs. TechPowerUp. 2026. 202URL: https://www.techpowerup.com/gpu-specs/geforce-gtx-1080-ti.c2877 (дата обращения: 29.04.2026)

8. Nvidia Nemotron. Nvidia. 2026. URL: https://developer.nvidia.com/nemotron (дата обращения: 29.04.2026)

9. NVIDIA Tesla P40 Specs. TechPowerUp. 2026. URL: https://www.techpowerup.com/gpu-specs/tesla-p40.c2878 (дата обращения: 29.04.2026)