МЕТОДИЧЕСКИЙ РАЗБОР · 28.05.2026

Накладные расходы
на обработку русского языка

Миф про 1.5×

7 экспериментов, 8 токенайзеров, 5 академических работ. Развенчание популярной цифры и пять рычагов экономии для МСБ.

АВТОР
Илья Утов

КАТЕГОРИЯ
AI · Tokenization · FinOps

TELEGRAM
t.me/
gorilla_under_hood

КОНТЕКСТ

Откуда берётся оверхед

Все большие модели разбивают текст на токены, атомарные единицы, которыми оперируют. Слово «tokenization» в новом токенайзере GPT-4o это один токен. Слово «эффективность» это два или три, в зависимости от поколения. Слово «несовершеннолетний» может уйти в четыре куска.

Причина простая. BPE-словари обучены на корпусах, где английский занимает 80-90% объёма. Любые подслова, частотные в английском, получают свой отдельный токен. Русские суффиксы, окончания, приставки собираются из тех «обрезков», что остались. Получается, что русское слово в среднем разбивается на 1.5-2 токена, тогда как английское на 1.1-1.3.

Это и есть «накладные расходы на обработку русского языка». Но он не одинаков везде. Я нарвался на минимум четыре фактора, которые меняют его в разы:

- Поколение токенайзера (GPT-3.5 vs GPT-4o, разница вдвое)
- Жанр текста (юридический канцелярит vs идиоматический чат, разница в 30 раз по проценту оверхеда)
- Размер промпта (на маленьких фрагментах эффект завышен из-за «прогрева»)
- Стиль письма внутри одного языка (об этом отдельный раздел, и это самая полезная часть)

ЭКСПЕРИМЕНТ

01 Размер имеет значение

Первое, что я померил, это зависимость оверхеда от длины промпта. Взял статью Wikipedia «Искусственный интеллект» на двух языках (84 тысячи символов на английском, 78 тысяч на русском) и нарезал её префиксами разной длины. Считал на двух токенайзерах OpenAI: `cl100k_base` (GPT-3.5/4, аналог Claude 2 эры) и `o200k_base` (GPT-4o, GPT-4.1, ближайший публичный прокси для свежего Claude).

РАЗМЕР ПРОМПТА → ОВЕРХЕД

короткий фрагмент
200 символов

длинный промпт
от 2 000 символов · плато

+61%

+27%

оверхед на коротком — артефакт «прогрева» BPE, не реальный налог

падение к плато

стабильно от 2K до 78K, это и есть реальный оверхед

Размер фрагмента	o200k ratio RU/EN	cl100k ratio RU/EN
200 символов	1.61× (+61%)	2.76× (+176%)
500 символов	1.48× (+48%)	2.68× (+168%)
2 000 символов	1.26× (+26%)	2.41× (+141%)
5 000 символов	1.27× (+27%)	2.49× (+149%)
20 000 символов	1.28× (+28%)	2.42× (+142%)
78 000 символов	1.27× (+27%)	2.31× (+131%)

На фрагменте в 200 символов русский в o200k выглядит на 61% тяжелее. На полной статье уже только 27%. Эффект выходит на плато к 2000 символов и дальше не меняется. Причина прозаична. Первые токены любого языка несут на себе всякую служебную нагрузку: заглавные, начала предложений, ведущие пробелы. На коротком тексте их доля раздута, на длинном растворена.

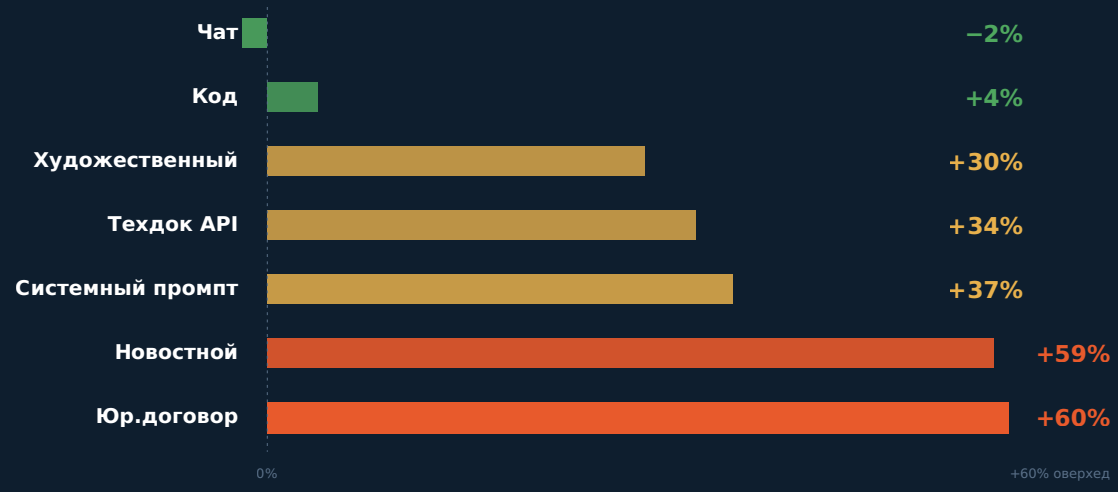
Вывод первый. Если кто-то меряет накладные расходы на обработку русского языка на одной фразе, верьте максимум 50%. Реальные системные промпты живут в режиме длинного текста, и там цифра в районе 25-30%.

ЭКСПЕРИМЕНТ

02 Жанровый срез

Самое интересное началось, когда я отказался от «общего русского» и стал мерить по жанрам. Семь параллельных пар: одна и та же мысль, выраженная идиоматически на каждом языке. Замер на `o200k_base`.

ЖАНРЫ ПО ОВЕРХЕДУ · ОТ ЛЁГКИХ К ТЯЖЁЛЫМ



Жанр	EN токенов	RU токенов	Оверхед
Юридический договор	48	77	**+60%**
Новостной	56	89	+59%
Системный промпт (методический)	52	71	+37%
Техническая документация API	62	83	+34%
Художественный	53	69	+30%
Код с комментариями	77	80	**+4%**
Чат, разговорный	50	49	**−2%**

Размах от минус двух до плюс шестидесяти. На идиоматическом разговорном русском один и тот же смысл укладывается в **меньше** токенов, чем на английском. На юридическом договорном в полтора раза больше.

Почему? В русском работают две силы, тянущие в разные стороны.

Сила, которая помогает русскому: морфологическая компактность. Падежи вместо предлогов. Аспект глагола вместо вспомогательных конструкций. Можно опустить подлежащее, потому что лицо зашито в окончание. «Прочти это» это одно повеление, и оно покрывает английское «I want you to read this». Безличные конструкции типа «холодно» против «it is getting cold». Свободный порядок слов вместо обязательного SVO. На разговорном тексте русский уходит вперёд: на пятнадцати тестовых парах он короче английского в среднем на 9% по токенам и на 45% по числу слов.

Сила, которая мешает русскому: длинные слова разбиваются на больше подслов. Каждое русское слово в среднем 1.6× больше токенов, чем английское. Слова «модернизация», «функциональность», «методология» режутся на 3-4 куска. Эта сила побеждает на бюрократическом, научном и юридическом текстах, там, где много длинных абстрактных существительных.

Где какая сила перевешивает, тот и платит. Чат и код это паритет или минус. Юридический договор и бюрократия плюс пятьдесят-шестьдесят. Wikipedia и техдок плюс тридцать.

ДВЕ СИЛЫ ВНУТРИ РУССКОГО ЯЗЫКА

МОРФОЛОГИЯ
сжимает русский

Падежи вместо предлогов
«другу» = «to the friend»

Опускание подлежащего
«Прочти это» = «I want you to read this»

−45%
меньше слов на ту же мысль

ТОКЕНИЗАЦИЯ
раздувает русский

Длинные слова режутся на части
«методология» → 3-4 токена

English-centric обучение
русские подслова из «обрезков»

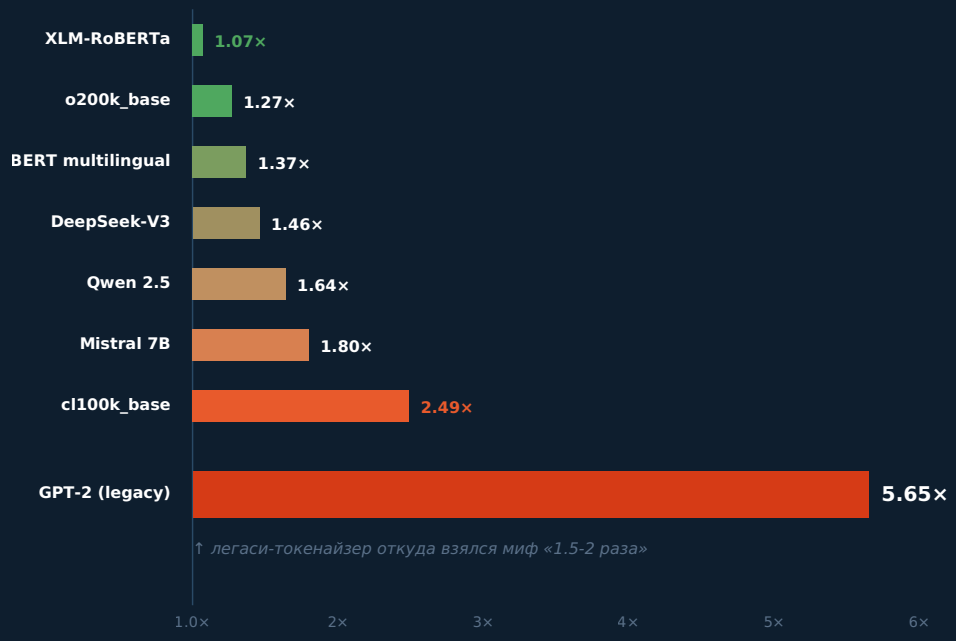
+1.6×
больше токенов на слово

ЭКСПЕРИМЕНТ

03 Спред токенайзеров

Третий замер. Один и тот же русский текст через разные токенайзеры. Брал фрагмент Wikipedia (5000 символов) и считал отношение RU/EN на восьми токенайзерах.

СПРЕД МЕЖДУ ТОКЕНАЙЗЕРАМИ · RATIO RU/EN НА ФОРМАЛЬНОМ ТЕКСТЕ



Токенайзер	Размер словаря	Ratio RU/EN на формальном	Ratio на чате
XLM-RoBERTa (мультиязычный)	250 000	**1.07×**	**0.79×**
o200k_base (GPT-4o, GPT-4.1)	200 000	1.27×	0.98×
BERT multilingual	110 000	1.37×	0.97×
DeepSeek-V3	128 000	1.46×	1.10×
Qwen 2.5	152 000	1.64×	1.19×
Mistral 7B	32 000	1.80×	1.14×
cl100k_base (GPT-3.5, GPT-4)	100 000	2.49×	1.58×
GPT-2 (legacy)	50 000	**5.65×**	3.12×

Спред пятикратный. Один и тот же русский текст можно посчитать как 1.07 умножить на английский (XLM-RoBERTa, специально обученный на 100 языках), или как 5.65 умножить на английский (старичок GPT-2, английский по построению).

Что важно для практики:

- Если вы работаете на современном Claude, GPT-4o, Gemini, Mistral, реальный коэффициент в диапазоне 1.3-1.7. «Полтора раза» это верхняя граница, не средняя.
- Если у вас застрял пайплайн на GPT-3.5 или старом фреймворке, да, там 2-3 раза. Это уже не из-за русского, а из-за выбора инструмента.
- Мультиязычные токенайзеры (XLM-RoBERTa, мульти-BERT) почти полностью убирают накладные расходы на обработку русского языка. Они не годятся для генерации, но для эмбеддингов или классификации, кейс есть.

Это совпадает с тем, что показала свежая работа Ovcharov 2026 «The Tokenizer Tax Across 25 European Languages». Там на болгарском (ближайший кириллический язык в их выборке) спред между топ-моделями составил 1.54×. Мой замер на русском дал близкий разброс между современными моделями (Mistral 1.80× против o200k 1.27× = 1.42×), и общий спред с включением legacy GPT-2, в пять раз. Направление совпадает с академией.

ЭКСПЕРИМЕНТ

04 Стиль письма важнее языка

Здесь начинается самый практически полезный кусок. Я взял один и тот же системный промпт в двух стилях, канцелярит и плотный, и замерил оба варианта на двух языках.

СТИЛЬ ВАЖНЕЕ ЯЗЫКА · ОДИН ПРОМПТ, ДВА РЕЖИМА

КАНЦЕЛЯРИТ

204 токенов · русский

190 токенов · английский

ПЛОТНЫЙ

83 токенов · русский

64 токенов · английский

РЫЧАГ СТИЛЯ

−59%

переписать в плотный стиль

РЫЧАГ ЯЗЫКА

−6%

сменить русский на английский

Канцелярит на русском (всё как любят писать в больших корпорациях):

> «Настоящим устанавливается, что ассистент должен осуществлять следующие функции. При получении любого запроса от пользователя необходимо в первую очередь произвести анализ контекста и определить, является ли запрос таковым, который требует обращения к внешним источникам информации…»

Тот же смысл, плотный стиль:

> «Ты ассистент. Правила. Перед ответом проверь, нужен ли поиск во внешних источниках. Нужен, ищи. В ответе источники и уверенность к каждому утверждению. Данных не хватает, скажи прямо. Не выдумывай. Без воды и повторов.»

Замер на `o200k_base`:

Вариант	Токенов	Слов
RU канцелярит	204	124
RU плотный	**83**	45
EN канцелярит	190	170
EN плотный	64	46

Канцелярит → плотный на русском режет токены в 2.5 раза (минус 59%). На английском минус 66%. Эффект больше, чем переход с русского на английский в любом стиле.

Сравните последний ряд. Русский плотный против английского плотного: 83 против 64, разница 30%. Русский канцелярит против английского плотного: 204 против 64, разница **в три раза**. Здесь два фактора накапливаются: канцелярит сам по себе даёт ×3 поверх плотного стиля, плюс русский поверх английского. В итоге трёхкратный разрыв.

Главный вывод эксперимента: **переписать промпт из канцелярита в плотный стиль даёт в десять раз больше экономии, чем выбор языка**. Если ваш системный промпт похож на договор с провайдером, основная переплата сидит не в языке.

ЭКСПЕРИМЕНТ

05 Бизнес-калькуляция

Соберу всё в деньги. Цены Anthropic на май 2026: Sonnet 4.6 это три доллара за миллион входных токенов и пятнадцать за миллион выходных. Opus 4.6 это пять и двадцать пять (после ноябрьского снижения 2025-го). Haiku 4.5 это доллар и пять.

Пять типовых сценариев для среднего бизнеса. Допущение: средний оверхед русского +27% (формальный стиль, плато из эксперимента 1). Важная оговорка: Claude tokenizer проприетарный, я мерил через o200k_base как прокси, реальный Claude может давать ±5 процентных пунктов к этим цифрам.

Сценарий	Запросов в месяц	EN $/мес	RU $/мес	Доп. $/год
Чат-бот поддержки (Sonnet)	50 000	$450	$572	**$1 458**
RAG по документам (Sonnet)	20 000	$660	$838	$2 138
Агент с инструментами (Sonnet)	10 000	$270	$343	$875
Аналитика на длинных промптах (Opus)	2 000	$300	$381	$972
Массовая классификация (Haiku)	500 000	$525	$667	$1 701
Итого		**$2 205**	**$2 800**	**$7 144**

При типичном миксе операций МСБ переплачивает за русский язык примерно **семь тысяч долларов в год**. Около 643 тысяч рублей по курсу 90. Не катастрофа, но и не пыль. Если оптимизировать стиль с +27% до +10%, экономия составит 13% от текущих расходов на русские промпты. Порядка 85 тысяч рублей в год на тех же объёмах.

При больших объёмах цифры пропорционально растут. Чат-бот с миллионом запросов в месяц вместо пятидесяти тысяч это уже не семь, а сто сорок тысяч долларов разницы в год.

Учтите: цифра +27% это среднее по плато. Для конкретного сценария реальный оверхед зависит от жанра (эксперимент 02): чат-бот может попасть ближе к нулю, аналитика на формальных промптах, к +37%. Считайте по своему профилю запросов, а не по среднему.

НАУКА

Что говорит наука

Чтобы не выглядеть как блогер с самописными графиками, перепроверил себя по академическим работам последних трёх лет.

Petrov, Malartic et al. 2023 «Language Model Tokenizers Introduce Unfairness Between Languages» (NeurIPS, arxiv 2305.15425). Измерили семнадцать токенайзеров на параллельном корпусе FLORES-200. Спред между языками до пятнадцати раз. Для болгарского (Cyrillic-прокси) ChatGPT тратит в 2.5 раза больше токенов, чем для английского. Но это на cl100k, легаси-токенайзер.

Ahia, Kumar et al. 2023 «Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models» (EMNLP). Зафиксировали динамику. Для русского fertility (среднее число токенов на слово) упала с 5.16 в 2021-м до 2.74 в 2023-м и продолжает падать. Это и есть та траектория, которая обнулила «миф 1.5-2 раза».

Tikhomirov, Chernyshev 2023 (МГУ, IEEE, arxiv 2312.02598) «Impact of Tokenization on LLaMa Russian Adaptation». Подменили словарь Llama на русско-специфичный: ускорение инференса до 60%, fine-tuning на 35%. Для тех, кому важно, техническая возможность убрать оверхед почти полностью существует, но требует переобучения модели.

Ali, Schubotz et al. 2024 «Tokenizer Choice For LLM Training: Negligible or Crucial?» (NAACL Findings, arxiv 2310.08754). English-центричный токенайзер на мультиязычной задаче добавляет до 68% оверхеда обучения. Мультиязычный требует словарь в три раза больше.

Ovcharov 2026 «The Tokenizer Tax Across 25 European Languages» (arxiv 2605.24718, май 2026). Самые свежие fertility-числа на современных моделях. Для болгарского: Gemma 2 это 2.35, GPT-4o это 2.45, Mistral Large

3 это 2.50, DeepSeek V3 это 3.07, Qwen3 32B это 3.62. Разница в 1.54 раза между лучшим и худшим современным токенайзером для одного и того же языка.

Вывод академии совпадает с моим домашним замером в трёх местах:

01 Цифра «полтора-два раза» это устаревшая оценка, она работала в эру GPT-3.5.

02 На современных моделях оверхед русского порядка 25-50%.

03 Выбор токенайзера влияет больше, чем выбор языка внутри одной модели.

РАЗДЕЛ

Где я честно расставил знаки вопроса

Anthropic не публикует свой токенайзер локально. Все мои числа для Claude это оценка через `o200k_base` как прокси (это публично доступный токенайзер OpenAI, близкий по поведению). Реальный Claude tokenizer может отличаться, точное измерение требует обращения к Anthropic API через метод `count_tokens` на конкретных production-промптах. Для оценки порядка величины при планировании бюджета прокси достаточен, для точного P&L, нужен свой замер.

Также я не делал статистики на большом корпусе. Семь жанровых пар это иллюстрация диапазона, не строгое распределение. Если у вас на руках свой production-промпт, лучший способ узнать ваши конкретные цифры это прогнать его через `count_tokens` обеих моделей и сравнить с эквивалентом на английском.

И ещё одно. Я подбирал пары руками. На пятнадцати чатовых парах русский оказался в среднем короче на 9%. На сто пятидесяти, возможно, картина была бы скромнее. Но направление эффекта (на идиоматическом разговорном русский в худшем случае паритет с английским) устоит.

ПРАКТИКА

Что с этим делать: пять рычагов экономии по убыванию эффекта

Если ваш AI-бюджет на русском промпте стал заметной строкой в P&L, вот порядок действий по силе эффекта. Я расположил их от самого мощного к самому слабому. Смена языка четвёртый из пяти.

Рычаг 1. Переписать промпты из канцелярита в плотный стиль.

Эффект: минус 40-60% токенов на самом промпте.

Что делать: убрать «настоящим устанавливается», «в случае необходимости», «следует осуществить». Заменить на императивы: «делай», «не делай», «проверь». Списки без вводных. Глаголы вместо отглагольных существительных. Этот рычаг не стоит ничего, кроме нескольких часов вдумчивой правки.

Рычаг 2. Включить prompt caching.

Эффект: минус 50-90% входных токенов при повторных запросах с одинаковым префиксом.

Что делать: у Anthropic кеширование работает с системными промптами. Заплатите один раз за full read, дальше префикс читается из кеша по сниженной ставке. На любом сценарии с длинным системным промптом и большим числом запросов это первое, что окупится.

Рычаг 3. Уменьшить контекст: убрать дубли, сжать историю, использовать summary-окно.

Эффект: минус 20-50%.

Что делать: ревизия того, что вообще летит в промпт. Старая история, повторяющиеся инструкции, забытые куски документации это всё чистый шум для модели и прямой расход.

Рычаг 4. Смешанный язык: системный промпт на английском, общение с пользователем на русском.

Эффект: минус 15-25% на системной части.

Что делать: системный промпт можно перевести на английский, модель спокойно отвечает на русском. Это уже сложнее, нужно следить, чтобы стиль ответа не съезжал. Но если системный промпт у вас тяжёлый, выигрыш реальный.

Рычаг 5. Поменять модель на ту, где токенайзер дружелюбнее к русскому.

Эффект: минус 10-30%.

Что делать: проверить, что у вашего класса задач делает Gemma, GPT-4o, Mistral. Для эмбеддингов посмотреть на XLM-RoBERTa. Но это сложный рычаг: модель это не только токенайзер. Качество ответа важнее.

Чего делать НЕ стоит. Переписывать весь продукт на английский «потому что русский дорогой». Эффект 25-30% экономии в лучшем случае. Цена: потеря смысла, доверия русскоязычной аудитории, конверсии. Сначала четыре рычага выше, и только потом, если ничего не помогло, думать про язык.

ИТОГ

Финал

Миф «русский в полтора раза дороже» родился в эру GPT-3.5 и закрепился у тех, кто не пересчитывал цифры с тех пор. Для современных моделей реальный диапазон: от паритета (чат, код-микс) до плюс шестидесяти процентов (бюрократический канцелярит). Среднее по большому корпусу около 25-30%, и эта цифра ещё снижается с каждой новой версией токенайзера.

Но самое важное даже не это. Внутри одного языка разница между плотным стилем и канцеляритом составляет 50-60% по токенам. Это вдвое больше, чем эффект от смены языка. Если вы оптимизируете расходы на AI, начинайте не с переводчика. Начинайте с редактуры.

Замеры и сырые таблицы у меня в Telegram. Если что-то захочется проверить на своём промпте, пишите туда, прогоню.

Источники и методология. Замеры, `tiktoken` (cl100k_base, o200k_base) и `transformers` (Mistral, Qwen 2.5, DeepSeek V3, BERT multilingual, XLM-RoBERTa, GPT-2) на параллельных текстах Wikipedia («Artificial intelligence» / «Искусственный интеллект») плюс семь жанровых пар авторских. Цены Anthropic, официальная страница на май 2026. Сырые скрипты замеров вышлю по запросу в Telegram. Академические источники: Petrov et al. 2023 (NeurIPS, arxiv:2305.15425), Ahia et al. 2023 (EMNLP, arxiv:2305.13707), Tikhomirov & Chernyshev 2023 (IEEE, arxiv:2312.02598), Ali et al. 2024 (NAACL, arxiv:2310.08754), Ovcharov 2026 (arxiv:2605.24718).

КОНЕЦ

Об этом документе

ДОКУМЕНТ	Накладные расходы на обработку русского языка в токенах · миф про 1.5×
ВЕРСИЯ	1.0 · собрано 28 мая 2026
АВТОР	Илья Утов
ОБЪЁМ	~2 700 слов · 22 страницы A4
ЖАНР	методический разбор для МСБ и AI-практиков
МЕТОДОЛОГИЯ	7 эмпирических экспериментов на tiktoken и transformers, параллельный корпус Wikipedia
ИСТОЧНИКИ	5 peer-reviewed работ: Petrov 2023 (NeurIPS), Ahia 2023 (EMNLP), Tikhomirov 2023 (IEEE), Ali 2024 (NAACL), Ovcharov 2026
ШЕРИНГ	будем рады распространению, ссылка на источник — приятно

ОБСУДИТЬ

У меня живой Telegram-канал

Сырые таблицы замеров, скрипты, разборы. Если есть свой production-промпт и хочется померить — присылайте, прогоню через все 8 токенайзеров и покажу диапазон.

t.me/gorilla_under_hood

НАВЕДИТЕ КАМЕРУ