ISSN : 2311-5122

7universum.com

UNIVERSUM: ТЕХНИЧЕСКИЕ НАУКИ

9(150)

НАУЧНЫЙ ЖУРНАЛ

...сандрович Попо́в (1821, Санкт-Петербург — ... Петербург) — русский флотоводец, ...роитель, адмирал (1891 год). Его имя носит ... Попова рядом с Владивостоком. Родился 22 сентября (4 октября) 1821 года в Санкт-Петербурге в семье известного кораблестроителя, управляющего Охтинской верфью Александра Андреевича Попова.

С 1830 года учился на морском отделение Александровского корпуса; в 1838 году окончил Морской кадетский корпус с присвоением чина мичмана и зачислением в 32-й флотский экипаж (Черноморский флот).

был командиром парохода «Метеор» (вспомогательного крейсера Черноморского флота). В 1853 году был командирован в Константинополь для сбора сведений о вооружении Босфора и близлежащих к нему укреплённых мест по Чёрному морю и Дунаю до Рущука. Участвовал в Крымской кампании. В 1853—1854 годах командовал пароходами «Эльборус», «Андия», «Турок», потопил шесть турецких транспортов; в сентябре 1854 года командовал в чине капитан-лейтенанта пароходом «Тамань», прорвался из блокированного Севастополя в Одессу и вернулся обратно с грузом для оборонявшихся.

Руководил установкой морских орудий на укреплениях Севастополя и артиллерийским снабжением. Руководил снаряжением двух брандеров. Руководил созданием бона для преграждения севастопольского рейда. За боевые заслуги в этот период — награждён двумя орденами.

Произведен в капитаны 1-го ранга; с 1855 — флигель-адъютант. В 1855 году был командирован с театра военных действий в Кронштадт, Выборг, Свеаборг и другие укреплённые порты побережья Балтийского моря. В период 1856—1858 годов был начальником штаба Кронштадтского порта.

В 1858—1861 гг. — командующий отрядом из двух корветов («Рында» и «Гридень») и клипера ... (2-й Амурский отряд), приведя его из Кронштадта в Японское море, затем ходил у берегов Японии, исследовал побережья русского Приморья (один из заливов — у 45-й параллели — был назван в честь корвета Попова «Рында»). В 1859 году ... командованием нанесли визит в порт Сан-Фр...

В 1861 году был произведен в контр-... назначением в Свиту[2]. ... Кораблестроительного и ... ведал переделкой парусн... Совершил плавание ... 1861 года был ... Отряда ... (Первая ...

205 ЛЕТ СО ДНЯ РОЖДЕНИЯ

21 сентября 2026 г.

ПОПОВ АНДРЕЙ АЛЕКСАНДРОВИЧ

1821-1898 гг.

РУССКИЙ ФЛОТОВОДЕЦ, КОРАБЛЕСТРОИТЕЛЬ, АДМИРАЛ

Г. МОСКВА

7universum.com
UNIVERSUM:
ТЕХНИЧЕСКИЕ НАУКИ

UNIVERSUM: ТЕХНИЧЕСКИЕ НАУКИ

Научный журнал

Издается ежемесячно с ноября 2013 года

Выпуск: 9(150)

Сентябрь 2026 г.

Часть 1

Новосибирск
2026

ББК 3
U55

U55 Universum: технические науки : электронный научный журнал. — № 9(150). — Часть 1. — Новосибирск : Изд-во «Юниверсум», 2026. — 88 с. — Текст : электронный. — URL: https://7universum.com/ru/tech/archive/category/9150 (дата публикации: 28.09.2026).

ISSN 2311-5122
DOI: 10.32743/UniTech.2026.150.9
Учредитель и издатель: ООО «Юниверсум»

Содержание

СТАТЬИ НА РУССКОМ ЯЗЫКЕ

ИНФОРМАЦИОННЫЕ И ЦИФРОВЫЕ ТЕХНОЛОГИИ

ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

Статья поступила в редакцию: 13.09.2026
Статья принята к публикации: 18.09.2026
Статья опубликована: 28.09.2026

УДК 330.564.2+004.9

DOI: 10.32743/UniTech.2026.150.9.23415

Исследование дифференциации доходов населения в Российской империи в 1905 году на базе программы Python

Борцов Александр Анатольевич
д-р техн. наук
гл. науч. сотрудник, НПО Квант
Российская Федерация, г. Москва
E-mail: laseroeo2@gmail.com

A study of income differentiation among the population of the Russian empire in 1905 based on a Python program

Bortsov Aleksandr A
Doctor of Technical Sciences,
Chief Researcher,
SFO 'Quantum'
Russian Federation, Moscow

Аннотация

В статье проведён количественный анализ дифференциации доходов населения Российской империи (РИ) в 1905 году. Актуальность работы определяется существенными расхождениями в существующих оценках уровня неравенства в дореволюционной России. Так, значения коэффициента Джини варьируются от 0,36 (Линдерта) до 0,55 (Новокмет). При этом для начала XIX века зафиксированы более высокие показатели — в диапазоне 0,65–0,67 (Корчмина). Значительные различия наблюдаются и в расчётах коэффициента

Библиографическое описание: Борцов А.А. Исследование дифференциации доходов населения в Российской империи в 1905 году на базе программы Python // Universum: технические науки : электрон. научн. журн. 2026. 9(150). URL: https://7universum.com/ru/tech/archive/item/23415

дифференциации KD, представленных Б.Н. Мироновым и Г.И. Ханиным. Причины этих расхождений кроются в интерпретации первичных данных, а также в неодинаковых подходах к учёту «самодеятельного» населения, фиксации доходов и определению численности состоятельных слоёв общества. *Целью работы является определение степени неравенства распределения доходов в РИ в 1905 году на основе кривой Лоренца, коэффициента Джини и децильного коэффициента дифференциации.* Для расчётов автором разработаны программы на языке Python, использующие данные о доходах и численности 30 социально-профессиональных групп, построенные на основе статистических отчётов РИ за 1905 год. Полученные результаты — коэффициент Джини G=0,51, коэффициент дифференциации KD=16,6 — подтверждают существенное расслоение по доходам. Аппроксимация кривой Лоренца ломаными линиями по методу наименьших квадратов выявила точку перелома на уровне ~98 % населения: последние 2 % сосредотачивают около 34 % всех доходов, а предельная доля дохода богатых превышает долю бедных в 25 раз. Распределение доходов по группам хорошо описывается законом Дагума (с параметрами c=1,18, d=1144, s=0,1). Результаты согласуются с оценками G Новокмета, Кирилова и KD Ханина и расходятся с заниженными, по мнению автора, данными Линдерта (G=0,36) и Миронова (KD=4,2–10,7).

Abstract

The article presents a quantitative analysis of income differentiation among the population of the Russian Empire (RE) in 1905. The relevance of the study is determined by the significant discrepancies in existing estimates of inequality levels in pre-revolutionary Russia. For instance, Gini coefficient values range from 0.36 (according to Linderth) to 0.55 (according to Novokmet). At the same time, higher figures were recorded for the early 19th century — in the range of 0.65–0.67 (Korchmina). Significant differences are also observed in the calculations of the differentiation coefficient KD presented by B.N. Mironov and G.I. Khanin. The reasons for these discrepancies lie in distortions in the primary data, in the differing approaches to accounting for the “self-employed” population, recording incomes, and determining the size of the affluent strata of society. The aim of the study is to determine the degree of inequality in income distribution in the RE in 1905 based on the Lorenz curve, the Gini coefficient, and the decile differentiation coefficient. For the calculations, the author developed programs in Python, using data on incomes and the size of 30 socio-professional groups; these data were compiled on the basis of statistical reports of the RE for 1905. The results obtained — the Gini coefficient G = 0.51 and the differentiation coefficient KD = 16.6 — confirm a substantial income stratification. Approximation of the Lorenz curve with broken lines using the least squares method revealed a turning point at approximately 98 % of the population: the last 2 % concentrate about 34 % of all incomes, and the marginal share of income among the rich exceeds that of the poor by a factor of 25. The income distribution across groups is well described by the Dagum distribution . The results are consistent with the estimates of G by Novokmet and Kirilov and of KD by Khanin, but diverge from what the author considers to be the underestimated data of Linderth (G = 0.36) and Mironov (KD = 4.2–10.7).

Ключевые слова: Python, неравенство доходов, Российская империя, 1905 год, коэффициент Джини, кривая Лоренца, децильный коэффициент дифференциации, распределение Дагума.

Keywords: Keywords: Python, income inequality, Russian Empire, Gini coefficient, Lorenz curve, decile coefficient, Dagum distribution, 1905.

Введение

Разные исследователи оценивают неравенство в дореволюционной России по-разному. Линдерт и др. получили коэффициент Джини около 0,36 для 1904 года [12], что соответствует умеренному неравенству. Новокмет с соавторами [10], опираясь на фискальные данные за 1905 год, получили более высокую оценку — около 0,55, но сами отметили возможные искажения из-за занижения крупных доходов и корректировки по коэффициентам Парето. Спорная оценка ван Зандена с соавторами на основе антропометрических данных за 1910 год [13] согласуется с результатами Линдерта [12]. Существует пока не подтверждённая академически точка зрения, что неравенство, достигавшее Джини 0,50 при отмене крепостного права, к 1890-м годам снизилось примерно до 0,40 и оставалось на этом уровне вплоть до 1910-х годов [4]. Однако при исследовании документов конкретных волостей исследователи отмечают, что статистические данные часто искажались, и это приводило к неверной оценке. Например, в работе [3] Кирилловым отмечено, что «действительный индекс Джини по с. Семилужному составлял не менее 0,438 (как в 1906 г.), а может быть, и 0,595 (как в 1904-м)». Другие работы по неравенству и дифференциации доходов населения РИ в 1905 г. с помощью оценки децильного коэффициента также не приводят к однозначной оценке, а зависят от принятого количества самодеятельного населения (49 млн чел. по оценкам Б.Н. Миронова [4] и более 60 млн чел. по оценкам Г.И. Ханина [6]) и их доходов. Б.Н. Миронов оценил децильный коэффициент для дореволюционной России (1905 г.) в диапазоне 4,2–10,7 [4], что сравнимо с Францией и Германией. Это существенно ниже оценок Г.И. Ханина (17,00–21,25) [6]. Занижение коэффициента Джини до 0,4 и ниже не согласуется с выводами, сделанными в работе [11]. В ней по первичным данным о доходах 7399 домохозяйств на материале Московской губернии 1811 года рассчитанный коэффициент Джини равен 0,65 (до налогов) и 0,67 (после налогов), при этом регрессивная налоговая система усугубляла высокое неравенство.

Целью исследования является анализ на базе разработанной программы Python дифференциации доходов населения РИ в 1905 г. Задачами исследования являются определение степени неравенства населения с помощью кривой Лоренца, коэффициента Джини и децильного коэффициента дифференциации доходов, а также аппроксимация распределений по известным законам.

Материалы и методы

Доходы населения и их количество в РИ за 1905 г., представленные в таблице 1, даны на основе анализа статистических отчётов РИ за 1905 год [10, 2, 4–7, 11]. Пользуясь статистическими методами оценки [8], развитыми для сложных систем, в работе используется математический аппарат анализа распределения доходов населения с аппроксимацией плотностей распределения.

Государство РИ было построено по иерархическому принципу. В отличие от чистой иерархической структуры (например, армии РИ) [1], в которой распределение доходов строго регламентировано, доходы населения РИ распределялись по смешанному принципу. С одной стороны, для государственных служащих, армии, полиции оклады были утверждены сверху, а доходы групп крестьян, рабочих, служащих частных предприятий, владельцев (земли, недвижимости, предприятий и т. п.) устанавливались экономическими связями благодаря рынку, посредством торговли. Упрощённая модель связей внутри различных слоёв населения РИ представлена в виде графа иерархического типа (рис. 1). Цветом выделены различные слои (крестьяне — светло-зелёный, рабочие — красный и др.).

Отличительной особенностью графа слоёв населения от графа чинов армии [8] является наличие распределения связей через рыночные структуры — торговлю и услуги. Связи становятся не однонаправленными, а двунаправленными. Тем не менее иерархические связи сохраняются внутри каждого слоя населения и между близкими группами по доходам. Расчёты производились с использованием разработанных автором программ на языке Python. Кривая Лоренца [9] определялась традиционно, как функция нормированных кумулятивных (нарастающих) доходов *Yk* от кумулятивного значения *Xk* количества членов в группе с составлением последовательности с нарастанием. Коэффициент Джини рассчитывался по формуле Брауна [1, 9].

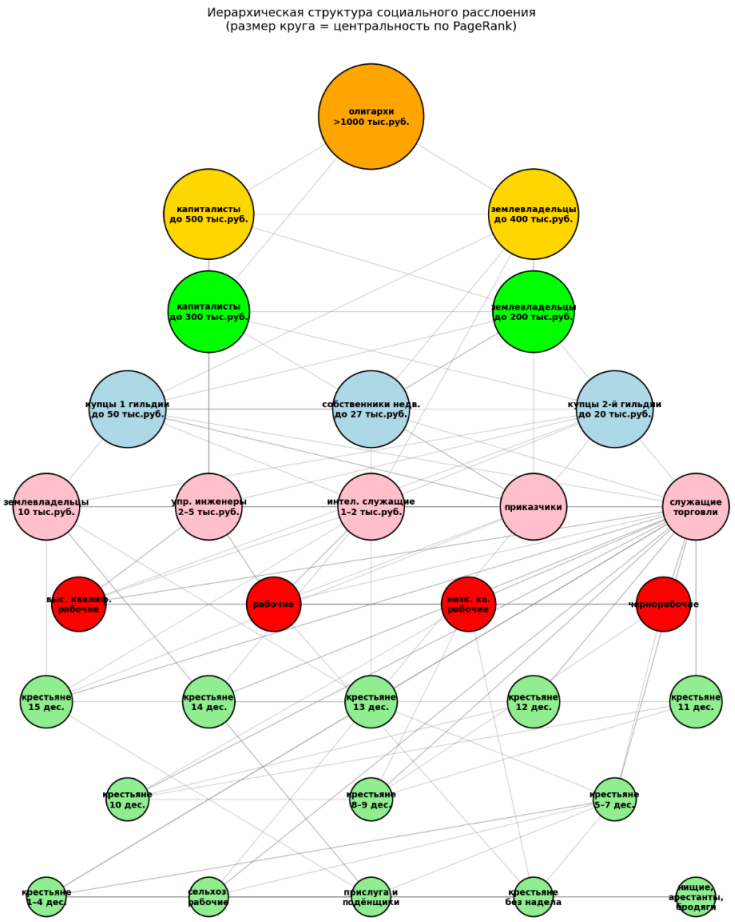

Рисунок 1. Структура графа связей групп активного населения РИ 1905 г.

В таблице 1 на основании исходных данных [10, 4, 6, 2, 5, 7] («Количество человек» и «Доход в рублях») приведены данные расчёта совокупного дохода группы как их произведение: «Доход группы в млн руб.» = «Количество чел.» ×× «Доход 1 чел. в руб.». Далее рассчитываются накопленное количество, нормированное на общее количество человек, и нормированный (на общую сумму дохода) накопленный доход. Было принято, что общее количество активного населения составляло 51 780 тыс. человек, а общий доход — 983,7 млн руб.

Таблица 1.

Месячные доходы населения РИ 1905 г.

Номер	Название группы	Количество в млн.чел.	Довольствие, руб.	Доход группы, млн. руб.	Кум. доля населения, ед.	Кум. доля дохода, ед.
1	нищие, арестанты, бродяги	0.101000	6	0.6	0.002	0.001
2	крестьяне без надела а	1.829000	6	11.3	0.037	0.012
3	крестьяне без надела б	2.461000	6	16.0	0.085	0.028
4	крестьяне без надела в	3.435000	7	24.0	0.151	0.053
5	прислуга и поденщики	3.332000	9	30.0	0.216	0.083
6	сельхоз.раб. 113 руб.	2.245000	9	21.1	0.259	0.105
7	крестьяне с наделом 1-4 десятин	4.713000	9	42.4	0.350	0.148
8	крестьяне с наделом 5-7 десятин	3.975000	10	39.8	0.427	0.188
9	чернорабочие	1.403000	13	18.2	0.454	0.207
10	рабочие с низк.квалификацией	5.223000	14	73.1	0.555	0.281
11	крестьяне с наделом 8-10 десятин	6.373000	15	95.6	0.679	0.378
12	крестьяне с наделом 11-12 десятин	5.186000	20	103.7	0.779	0.484
13	Крестьяне с наделом 13 десятин	4.680000	24	112.3	0.869	0.598
14	Крестьяне с наделом 14 десятин	1.545000	26	40.2	0.899	0.639
15	Крестьяне с наделом 15 десятин	1.449000	28	40.6	0.927	0.680
16	рабочие средней квалификации	1.478000	23	34.0	0.956	0.715
17	служащие торговли и услуг	1.275000	36	45.9	0.981	0.761
18	рабочие высокой квалификации	0.305000	65	19.8	0.986	0.781
19	приказчики в торговле и услугах	0.293000	70	20.5	0.992	0.802
20	с дох. в 1000-2000 р. в год	0.220000	120	26.4	0.996	0.829
21	с дох. в 2001-5000 р.в год	0.121000	250	30.2	0.999	0.860
22	с дох. в 5001-10000 руб. в год	0.037000	550	20.4	0.999	0.881
23	с дох. в 10001-20000 руб. в год	0.016000	993	15.9	1.000	0.897
24	с дох. 20001-27000 руб. в год	0.006000	1,700	10.2	1.000	0.907
25	с дох. 27001-50000 руб. в год	0.003000	2,500	7.5	1.000	0.915
26	с дох. 50001-200000 руб. в год	0.002150	4,167	9.0	1.000	0.924
27	с дох. 200001-300000 руб. в год	0.001150	16,667	19.2	1.000	0.943
28	с дох. 300001-500000 руб. в год	0.000650	25,000	16.2	1.000	0.960
29	с дох. 500001-1000000 руб. в год	0.000450	41,667	18.8	1.000	0.979
30	с дох. более 1000000 руб. в год	0.000250	83,333	20.8	1.000	1.000
ИТОГО	ВСЕГО	51.708650	-	983.7	1.000	1.000

На рис. 2 приведена гистограмма распределения месячных доходов по 30 социально-профессиональным группам населения Российской империи в 1905 году (данные второго столбца таблицы 1). Разрыв в доходах между самой обеспеченной 30-й группой и одной из наиболее многочисленных и наименее обеспеченных — 2-й группой («крестьяне без надела», средний доход — 6,2 рубля в месяц) — превышает 13 440 раз.

На рис. 3 отражена численность активного населения по группам в 1905 году. Распределение совокупных доходов по группам (рассчитанное как произведение месячного дохода на численность группы) представлено на рис. 4.

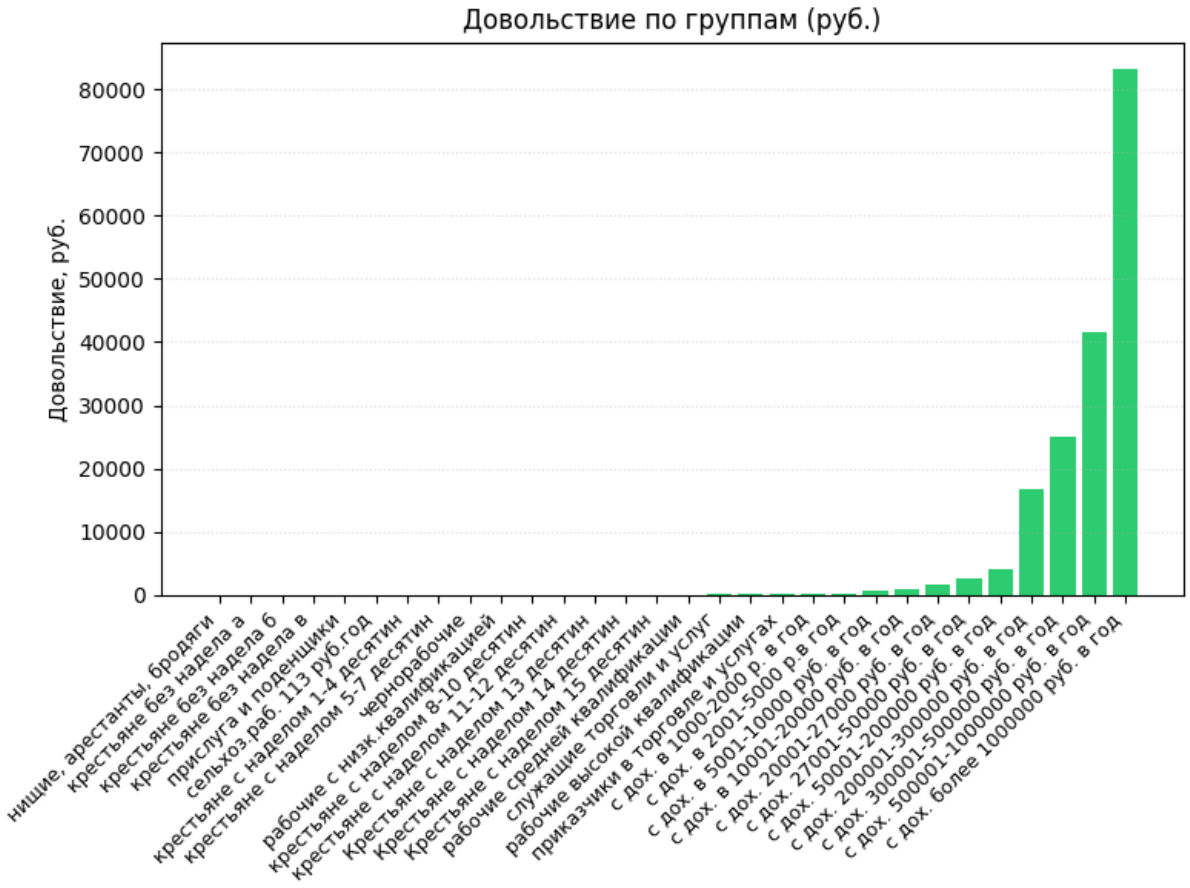

Рисунок 2. Довольствие(доход) в группе РИ в 1905 г.

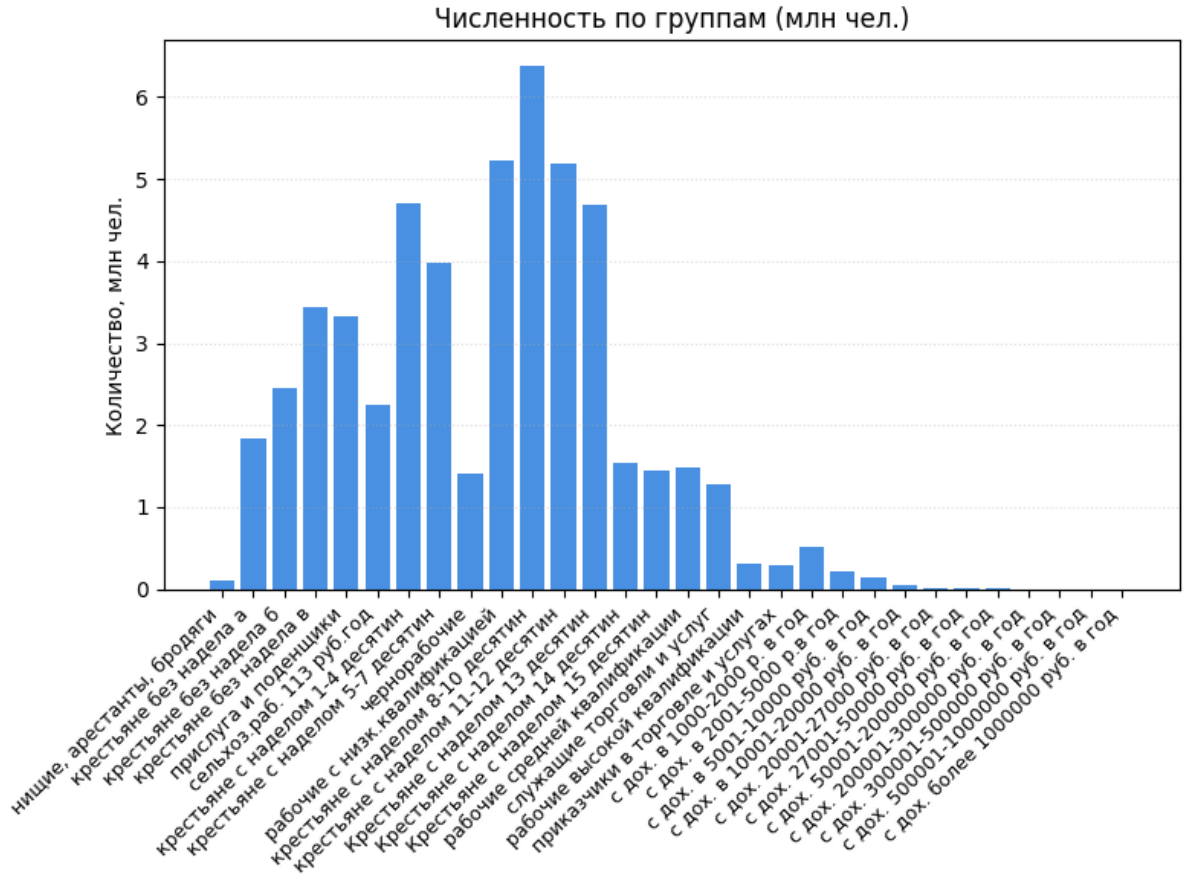

Рисунок 3. Количество активного населения по группам РИ в 1905 г.

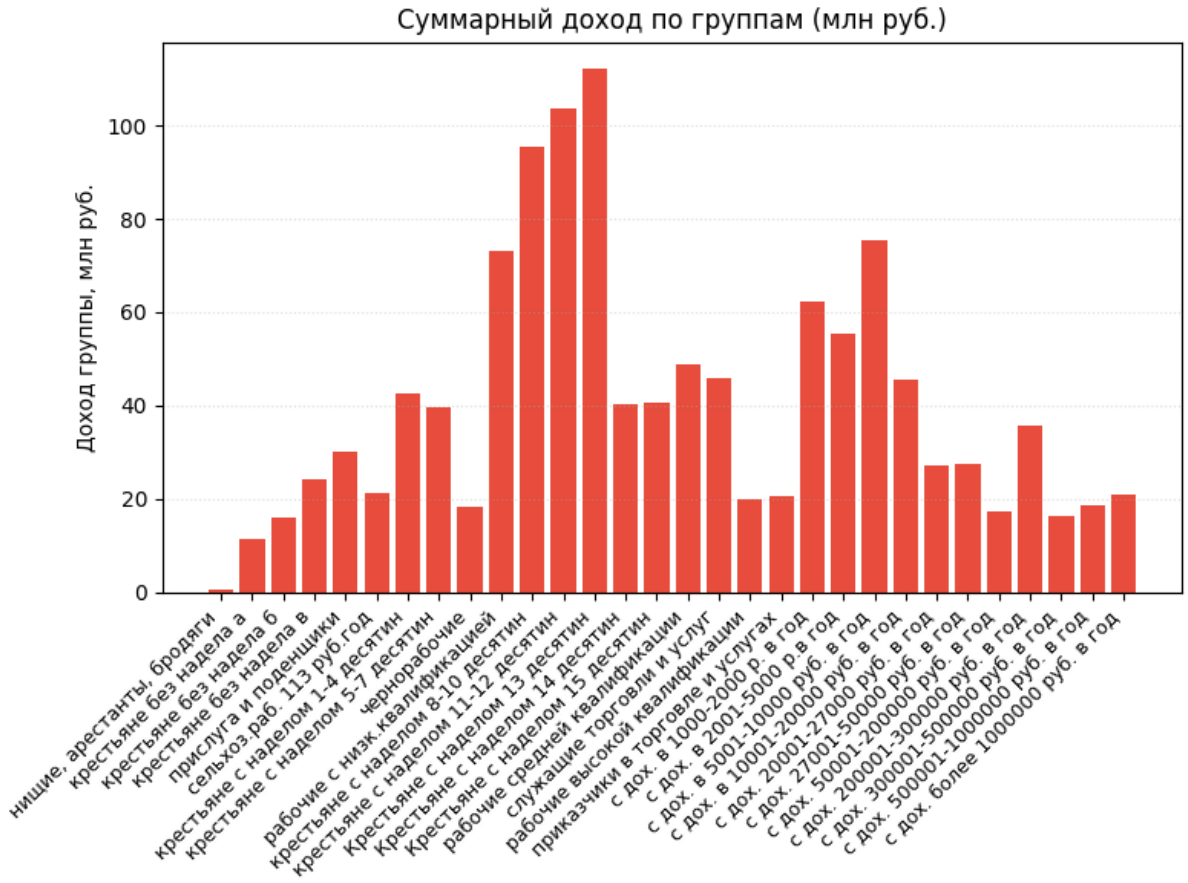

Рисунок 4. Суммарные доходы каждой группы в месяц РИ в 1905 г.

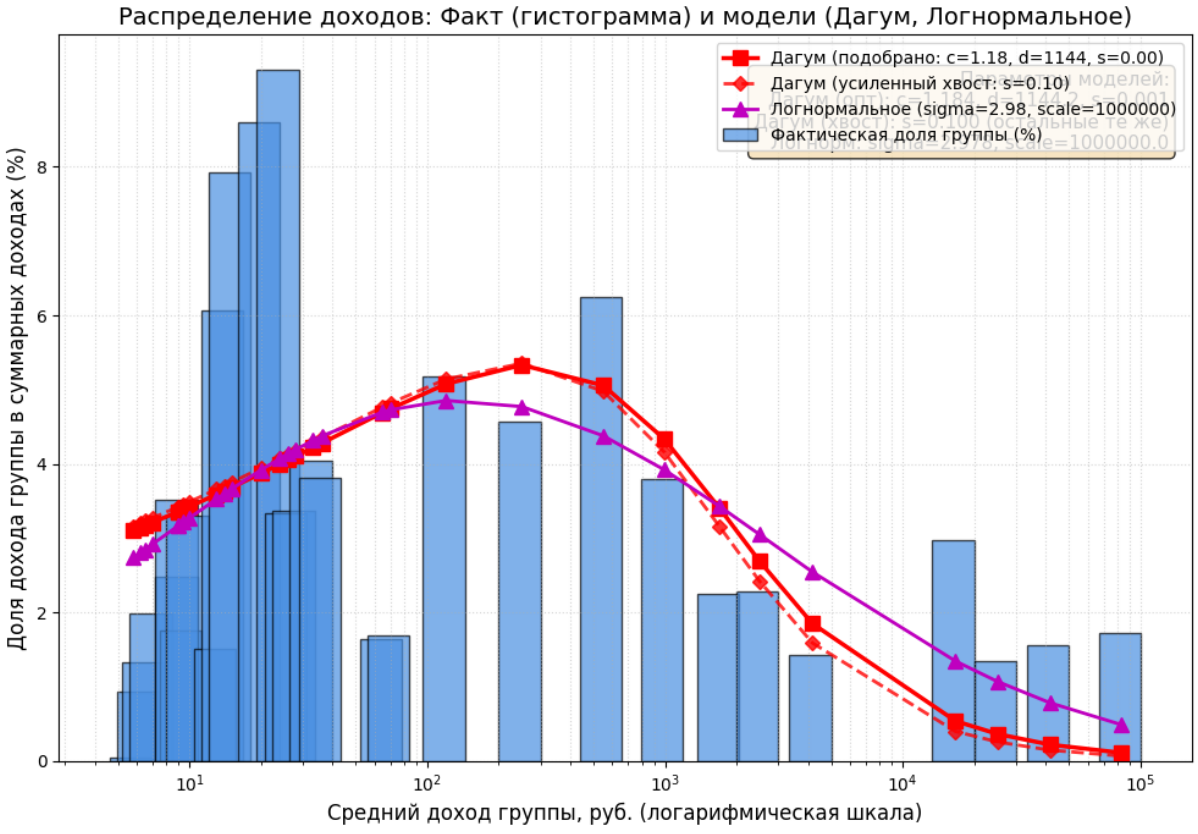

Рисунок 5. Суммарные доходы по группам в месяц РИ 1905 г.

На рис. 5 представлена гистограмма распределения доходов по группам населения: столбцы отражают величину доходов, по оси абсцисс отложена логарифмическая шкала среднего дохода группы, по оси ординат — доля дохода каждой группы в процентах. Наибольший объём доходов приходится на крестьянские группы с 11-й по 15-ю, члены которых владели земельными наделами размером от 8 до 13 десятин (1 десятина = 1,09 гектар или 10 900 м²). При этом совокупные доходы группы рабочих средней квалификации (16-я группа) сопоставимы с доходами групп крестьян, имевших надел в 14 десятин (14-я группа) и 15 десятин (15-я группа).

Плотность распределения доходов по группам достаточно точно описывается двумя моделями: распределением Дагума (с параметрами c=1,18, d=1144, s=p=0,1 — последний параметр характеризует «хвост» распределения) и логнормальным распределением (с параметрами σ(sigma)=2,98, scale=1 000 000).

На рис. 6 сплошная кривая отражает кривую Лоренца, построенную по данным таблицы 1, — она наглядно демонстрирует степень неравенства в распределении месячных доходов между различными группами активного населения. Согласно проведённым расчётам на основе указанных данных, коэффициент Джини составил G=**0,514**, а коэффициент дифференциации доходов — KD=**16,67**.

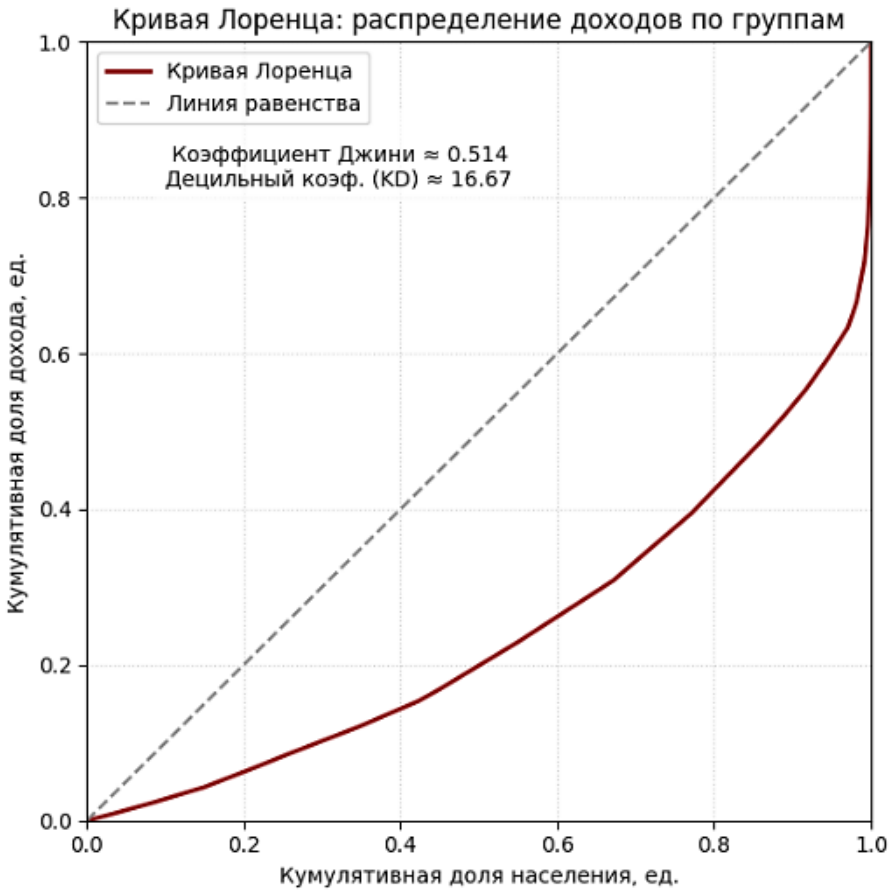

Рисунок 6. Кривая Лоренца доходов активного населения РИ 1905 г.

Проведено сопоставление кривых Лоренца, построенных двумя способами: на основе исходных данных (табл. 1) и с использованием распределения Дагума (с параметрами c=1,18, d=1144, s=0,1) — результаты представлены на рис. 5. Распределение Дагума демонстрирует

высокую аппроксимирующую способность при оценке доходного неравенства: погрешность расчёта кривой Лоренца по этой модели не превышает 1 % по сравнению с расчётом непосредственно по эмпирическим данным (табл. 1).

Достоверность исходной модели косвенно подтверждается тесной связью между распределением доходов гражданского населения и распределением довольствия чинов армии Российской империи в 1905 году. Коэффициент корреляции между соответствующими кривыми Лоренца составил R=0,985 (источник [8]), что указывает на высокую степень согласованности процессов иерархического и рыночного распределения доходов — ведь армейские чины являлись неотъемлемой частью общего населения РИ. Линейная регрессия, описывающая эту связь ($Y=AX+B$, где X — кумулятивное довольствие в армии, Y — кумулятивные доходы населения), характеризуется коэффициентами A=1,054 и B=0,013.

Для выделения групп с наиболее высокими доходами применена аппроксимация кривой Лоренца (рис. 6) ломаными линиями на базе метода наименьших квадратов — результаты отражены на рис. 7–9. Такой подход позволяет чётко обозначить сегменты распределения, в которых концентрируются максимальные доходы, а используемый статистический критерий помогает оценить точность прогнозных моделей.

При аппроксимации кривой Лоренца *двумя ломаными* линиями была определена точка излома x_0≈0,980, y_1≈0,66 (рис. 7). Это означает, что перелом в распределении доходов приходится примерно на 98 % населения — соответственно, оставшиеся 2 % получают 34 % всех доходов. При этом предельная доля дохода у наиболее обеспеченной группы превышает долю наименее обеспеченной в 25 раз.

При уточнении модели с помощью аппроксимации *тремя ломаными линиями* при сохранении первой точки излома x_0≈0,980, y_1≈0,66 выявлена вторая точка излома на базе метода наименьших квадратов — x_1≈0,9993, y_1≈0,86 (рис. 8). Таким образом, «второй перелом» в распределении доходов приходится примерно на 99,93 % населения — соответственно, оставшиеся 0,07 % населения получают 14 % всех доходов.

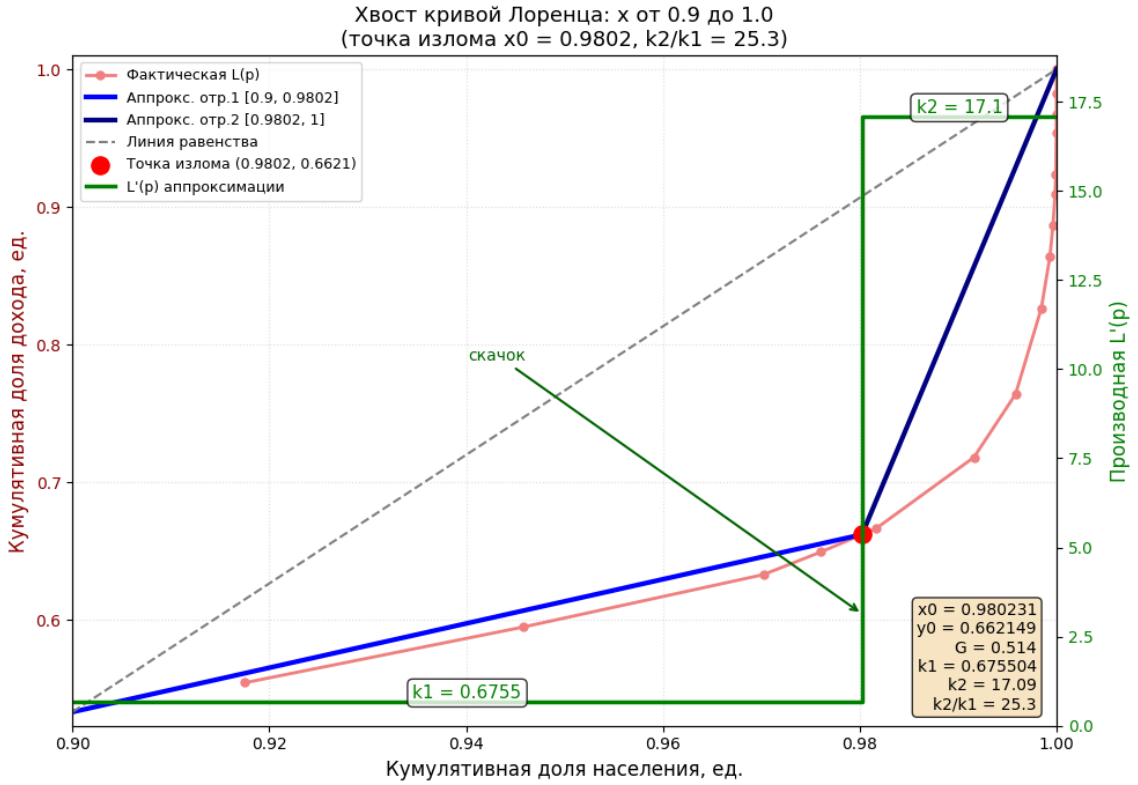

Рисунок 7. Аппроксимация кривой Лоренца двумя ломаными линиями

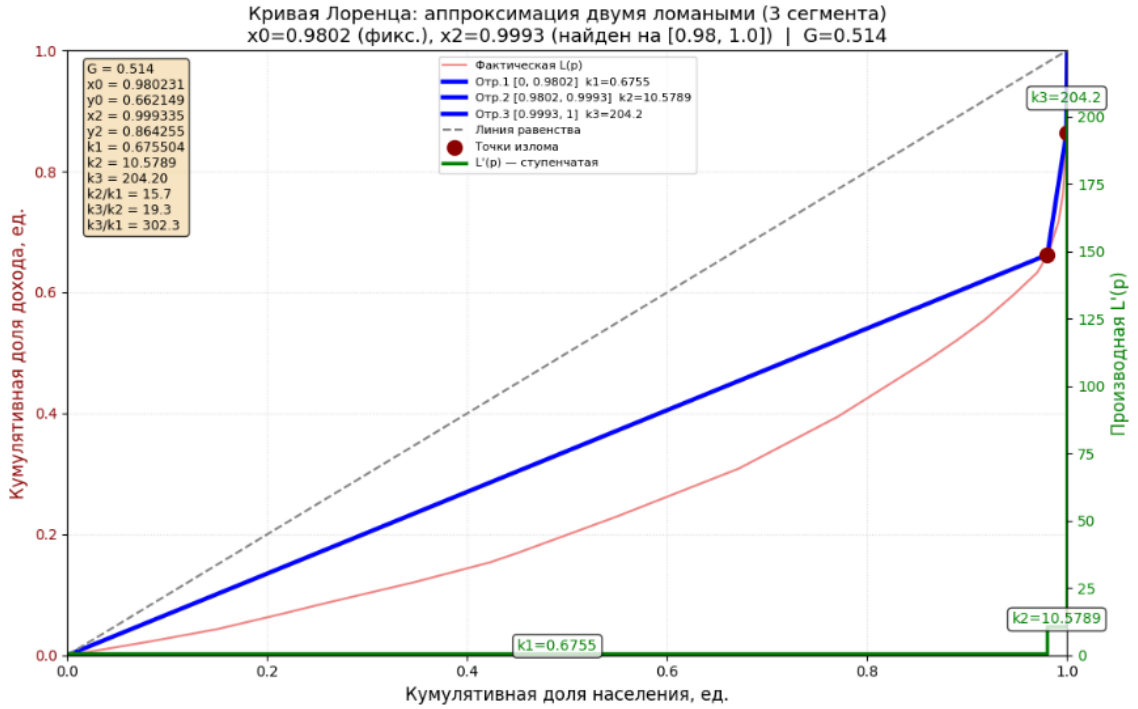

Рисунок 8. Аппроксимация 3-мя ломанными кривой Лоренца (интервал от 0,9955 до 1,0000) доходов активного населения РИ 1905 г.

Средние крутизны линии Лоренца на выделенных интервалах, рассчитанные методом наименьших квадратов, демонстрируют резкий контраст между сегментами распределения: на интервале (0; 0,98) средняя крутизна составляет $k1$=0,68, тогда как на интервале (0,98; 1,00) она резко возрастает до $k2$=10,6. Это наглядно отражает высокую концентрацию доходов в верхней части распределения.

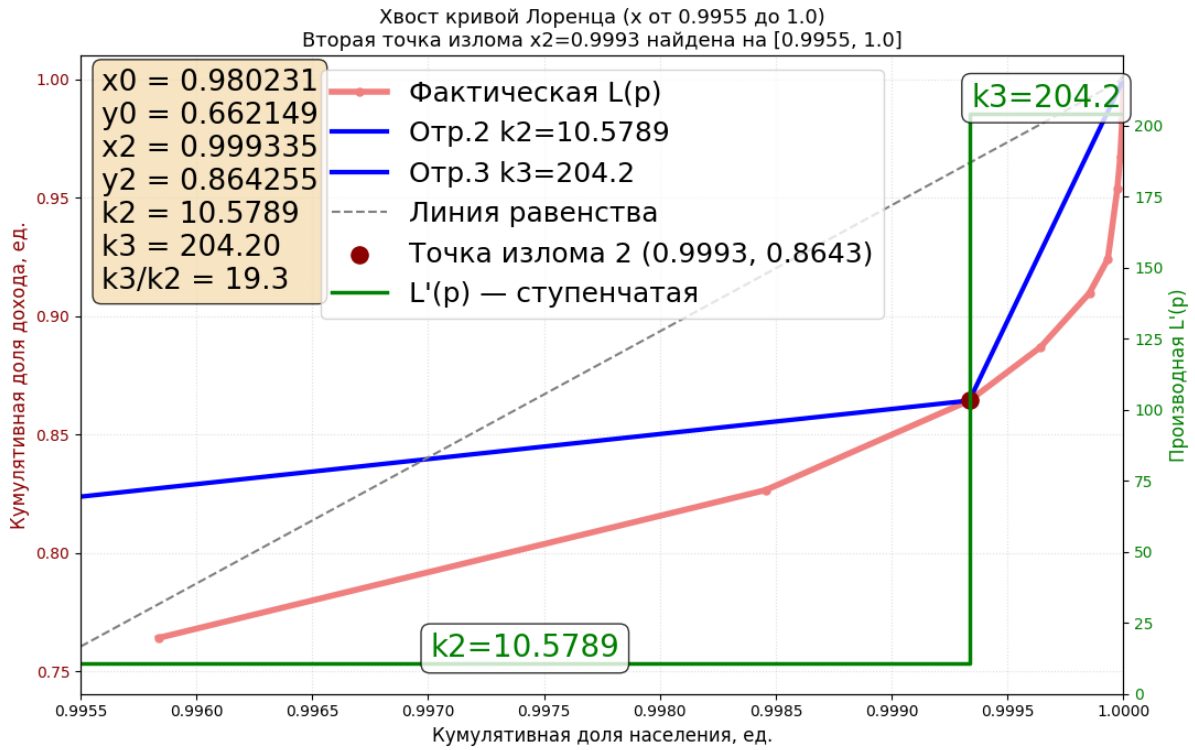

Рисунок 9. Аппроксимация 3-мя ломанными кривой Лоренца (интервал от 0,9955 до 1,0000) доходов активного населения РИ 1905 г.

Результаты и обсуждение

Точки излома, выявленные при аппроксимации кривой Лоренца ломаными линиями, наглядно выделяют группы с наиболее высокими доходами, а также демонстрируют динамику роста производных на интервалах между этими группами. Согласно принятым оценкам распределения доходов (табл. 1), коэффициент Джини составил G=0,51, а коэффициент дифференциации — KD=16,6. Полученные значения подтверждают наличие существенного социально-экономического расслоения и выраженной дифференциации доходов в Российской империи в 1905 году.

Средние крутизны роста доходов, рассчитанные методом аппроксимации кривой Лоренца, отчётливо показывают неравномерность распределения. При использовании двух ломаных линий получены следующие значения: на интервале (0; 0,98) средняя крутизна равна k1=0,68, а на интервале (0,98; 1,00) она резко возрастает до k2=10,6.

При уточнении модели с помощью трёх ломаных (с первой опорной точкой 0,98) картина становится ещё более детализированной: на интервале (0; 0,98) крутизна остаётся на уровне k1=0,68; на участке (0,98; 0,9993) она составляет k2=10,6; а на финальном отрезке (0,9993; 1,0000) достигает экстремально высокого значения k3=204,2 (рис. 4). Такой резкий рост крутизны на последнем интервале отражает концентрацию значительной доли доходов у крайне узкой элитарной группы населения.

Предложенный метод выделения элитных групп на основе средней крутизны роста доходов демонстрирует высокую эффективность. Вывод очевиден: именно самые богатые слои общества характеризуются наибольшими значениями крутизны, что свидетельствует о максимальной концентрации доходов в верхней части распределения. В случае РИ 1905

года это проявляется особенно наглядно: на финальном участке кривой Лоренца крутизна достигает экстремального значения k3=204, фиксируя, что крайне узкая группа населения аккумулирует значительную долю совокупного дохода.

Такая высокая крутизна отражает не просто разницу в уровне благосостояния, а принципиально иной масштаб прироста доходов у элиты по сравнению с остальными слоями. На интервале (0,9993; 1,0000) даже минимальное приращение доли населения сопровождается колоссальным приростом доли дохода — это и есть математическое выражение гиперконцентрации ресурсов.

Как известно, благодаря кумулятивному эффекту концентрация у элитарного слоя всё большего дохода с течением времени может привести к критическому значению неравенства. Механизм этого эффекта строится на положительной обратной связи: накопленный капитал позволяет получать ещё больший доход — через владение землёй, недвижимостью, предприятиями, доступ к более выгодным финансовым инструментам и привилегированным рыночным позициям. Доходы элиты не просто выше — они воспроизводятся и приумножаются быстрее, чем доходы остальных групп, поскольку сами становятся источником новых доходов.

В условиях Российской империи начала XX века этот кумулятивный эффект усиливался институциональными факторами: сохранением сословных привилегий, ограниченным доступом широких слоёв к образованию и мобильности, а также особенностями налоговой системы, которая зачастую носила регрессивный характер. В результате даже при умеренном росте общего благосостояния разрыв между верхними и остальными слоями не сокращался, а, напротив, мог нарастать — ведь темпы прироста доходов у богатых опережали средние темпы экономического роста.

Критическое значение неравенства в таком контексте — это не просто высокий коэффициент Джини, а состояние, при котором концентрация доходов начинает подрывать социальную устойчивость: снижается экономическая мобильность, нарастает ощущение несправедливости, падает доверие к институтам. Исторически такие состояния нередко становились источником социальной напряжённости и политических кризисов. В этом смысле количественные оценки неравенства (вроде G=0,51 и KD=16,6) — не просто статистические показатели, а индикаторы глубинных структурных диспропорций, способных определять траекторию развития общества в долгосрочной перспективе.

Заключение

В результате численного расчёта на основе исторических данных о доходах населения Российской империи в 1905 г. были получены кривая Лоренца, коэффициент Джини G=0,51 и коэффициент дифференциации KD=16,6. Эти значения, рассчитанные по оценочным исходным данным, хорошо согласуются с результатами работ Новокмета, Кириллова и Ханина [10, 3, 6] и подтверждают сильную дифференциацию доходов населения Российской империи в 1905 г. В то же время они существенно превышают низкие оценки G и KD , приведённые в работах Линдерта (G=0,36), ван Зандена и Миронова (KD=от 4,2 до 10,7) [12, 13, 4].

Наличие в дореволюционной Российской империи глубокого неравенства ($G \geq 0{,}51$) косвенно подтверждается рядом независимых свидетельств. Во-первых, период с 1891 по 1912 г. охватил серия очагов массового голода, среди которых наиболее разрушительными были голод 1891–1892 гг., охвативший около 30 млн. человек в Центрально-Черноземном и Поволжском регионах, и продовольственный кризис 1897–1898 гг., поразивший до 18 губерний. Регулярность этих катастроф указывает на хроническую неспособность значительной части населения — прежде всего крестьянства — обеспечивать минимальный уровень потребления даже в условиях роста экономики в целом. Во-вторых, хотя крепостное право было отменено в 1861 г., земельная реформа так и не была доведена до завершения: «земельный вопрос» остался нерешённым, малоземелье и чересполосица сохранялись, а выкупные платежи тяжким бременем ложились на крестьянские хозяйства вплоть до их отмены в 1907 г. Это означает, что институциональные предпосылки высокого неравенства, унаследованные от крепостнической эпохи, продолжали действовать и в начале XX в.

Ошибка (до 9 %) в определении коэффициента Джини в настоящей работе обусловлена в основном неопределённостью данных о численности высших социальных групп (с 24-й по 30-ю) и может быть уточнена в сторону увеличения при более детальном анализе состава этих групп.

Распределение Дагума даёт хорошую аппроксимацию и эффективно при оценке неравенства доходов населения: ошибка расчёта кривой Лоренца составляет не более 1 % по сравнению с расчётом непосредственно по исходным данным (табл. 1).

Список литературы

1. Борцов А.А. Анализ на базе программы Phyton дифференциации довольствия нижних чинов, офицеров и генералов в армии Российской империи в 1905 году / А.А. Борцов // Universum: технические науки: электрон. научн. журн. – 2026. – № 8 (149). – DOI: 10.32743/UniTech.2026.149.8.23253.
2. Довель М. Узкие полосы // Тверские губернские ведомости. – 1906. – № 29.
3. Кириллов А.К., Резникова М.А. Неравенство доходов крестьян пригородной волости по данным налоговыхраскладок: Семилужная волость Томского уезда в начале XX в // Исторический журнал: научные исследования. – 2023.
4. Миронов Б.Н. «Благосостояние населения и революции в имперской России: XVIII – начало XX века» монография (2010, 2-е изд. 2012).
5. Статистика землевладения 1905 г.: Свод данных по 50-ти губерниям Европейской России. СПб. – СПб. – 1907.
6. Ханин Г.И. Какова была социальная дифференциация в дореволюционной России? – Идеи и идеалы № 2(4), т. 1, 2010, с 64–72.
7. Щербина Ф.А. Крестьянские бюджеты / Ф.А. Щербина. – Воронеж: Типография В.И. Исаева, 1900. – 730 с.
8. Bortsov A.A., Il'in Yu. B., Smolskiy S.M. Laser Optoelectronic Oscillators. – Cham: Springer, 2020. – XXXV, 522 P.(Springer Series in Optical Sciences; Vol. 232).

9. Brown M.C. «Using Gini-style indices to evaluate the spatial patterns of health practitioners: theoretical considerations and an application based on Alberta data» // Social Science & Medicine, –1994. – 1994. – T. 38, № 9.
10. Filip Novokmet, Thomas Piketty, Gabriel Zucman. «From Soviets to Oligarchs: Inequality and Property in Russia 1905–2016.» The Journal of Economic Inequality, 2018, vol. 16, no. 2, P. 189–223.
11. Korchmina M., Malinowski. Korchmina, Malinowski. «How Extractive Was Russian Serfdom? Income Inequality in Moscow Province in the Early 19th Century.» EHES Working Paper // EHES Working Paper. – 2024. – № 266 (2024).
12. Peter H. Lindert, Steven Nafziger. «Russian Inequality on the Eve of Revolution.» The Journal of Economic History, 2014, vol. 74, no. 3, P. 767–798.
13. Van Zanden J.L. How Was Life? Global Well-being since 1820 / J.L. van Zanden, J. Baten, M. Mira d'Ercole, A. Rijpma, C. Smith, M. Timmer. – Paris: OECD Publishing, 2014. – 272 P.

References

1. Bortsov A.A. Analysis of Differentiation of Rations of Lower Ranks, Officers, and Generals in the Russian Imperial Army in 1905 Based on a Python Program // Universum: Technical Sciences: Electronic Scientific Journal. – 2026. – No. 8 (149).
2. Dovel M. Narrow Strips // Tver Provincial Gazette. – 1906. – No. 29.
3. Kirillov A.K., Reznikova M.A. Income Inequality of Peasants in a Suburban Volost According to Tax Assessment Records: Semiluzhnaya Volost, Tomsk Uyezd, Early 20th Century // Historical Journal: Scientific Research. – 2023. – No. 6.
4. Mironov B.N. The Well-Being of the Population and Revolutions in Imperial Russia: 18th – Early 20th Century. – Monograph (2010, 2nd ed. 2012).
5. Statistics of Land Ownership in 1905: Summary of Data for 50 Provinces of European Russia. – St. Petersburg. – St. Petersburg. – 1907.
6. Khanin G.I. What Was the Social Differentiation in Pre-Revolutionary Russia? // Ideas and Ideals. – 2010. – Vol. 1, No. 2 (4).
7. Shcherbina F.A. Peasant Budgets. – Voronezh: Typography of V.I. Isaev, 1900.
8. Bortsov A.A., Il'in Yu.B., Smolskiy S.M. Laser Optoelectronic Oscillators. – Cham: Springer, 2020. – XXXV, 522 P. – (Springer Series in Optical Sciences; Vol. 232).
9. Brown M.C. «Using Gini-style indices to evaluate the spatial patterns of health practitioners: theoretical considerations and an application based on Alberta data» // Social Science & Medicine. – 1994. – Vol. 38, No. 9.
10. Filip Novokmet, Thomas Piketty, Gabriel Zucman. «From Soviets to Oligarchs: Inequality and Property in Russia 1905–2016.» The Journal of Economic Inequality, 2018, vol. 16, no. 2, P. 189–223.
11. Korchmina E., Malinowski M. «How Extractive Was Russian Serfdom? Income Inequality in Moscow Province in the Early 19th Century.» EHES Working Paper No. 266 (2024).

12. Peter H. Lindert, Steven Nafziger. «Russian Inequality on the Eve of Revolution.» The Journal of Economic History, 2014, vol. 74, no. 3, P. 767–798.
13. van Zanden J.L., Baten J., Mira d'Ercole M., Rijpma A., Smith C., Timmer M. Jan Luiten van Zanden, Joerg Baten, Marco Mira d'Ercole, Auke Rijpma, Constance Smith, Marcel Timmer (eds.). «How Was Life? Global Well-being since 1820.» OECD Publishing. – Paris: OECD Publishing, 2014.

Статья поступила в редакцию: 13.08.2026
Статья принята к публикации: 15.09.2026
Статья опубликована: 28.09.2026

УДК 004.7

Топологии и защита информации

Сухачёв Никита Алексеевич
студент кафедра информационной безопасности Технологический университет имени А.А. Леонова
Российская Федерация, г. Королёв
E-mail: brawl.sukh@mail.ru

Теребина Анна Александровна
Студентка кафедра информационной безопасности Технологический университет имени А.А. Леонова
Российская Федерация, г. Королёв
E-mail: ANNA.TEREBINA@yandex.ru

Topologies and information security

Sukachev Nikita Alekseevich
student, Department of Information Security Leonov Technological University
Russia, Korolev

Terebina Anna Aleksandrovna
student, Department of Information Security Leonov Technological University
Russia, Korolev

Аннотация

В современном мире, где информация стала одним из самых ценных ресурсов, её защита — не просто техническая задача, а вопрос выживания бизнеса и доверия пользователей. Каждый день мы слышим о новых кибератаках, утечках данных и сбоях в работе критически важных систем — и всё это напоминает: уязвимость может скрываться не только в коде или настройках, но и в самой архитектуре сети. Именно поэтому вопрос выбора правильной топологии нельзя считать второстепенным: от того, как соединены узлы, напрямую зависит, насколько надёжно будут храниться и передаваться данные. В данной работе рассматривается влияние топологии компьютерной сети на уровень защищённости информации. Проведён сравнительный анализ четырёх базовых топологий («шина», «кольцо», «звезда», «смешанная») по таким критериям, как конфиденциальность, целостность, доступность, устойчивость к отказам, масштабируемость и сложность обнаружения вторжений. Для количественной оценки уязвимостей предложена авторская модель расчёта коэффициента защиты информации (КЗИ), основанная на методологии

Библиографическое описание: Сухачёв Н.А., Теребина А.А. Топологии и защита информации // Universum: технические науки : электрон. научн. журн. 2026. 9(150). URL: https://7universum.com/ru/tech/archive/item/23413

ФСТЭК. Результаты расчёта показывают, что наилучшую защищённость обеспечивает смешанная топология (КЗИ = 0,14), а топология «звезда» (КЗИ = 0,34) является оптимальным выбором для большинства корпоративных сетей.

Abstract

In today's world, where information has become one of the most valuable resources, its protection is not merely a technical task but a matter of business survival and user trust. Every day we hear about new cyberattacks, data leaks, and failures in critical systems — all of which remind us that vulnerabilities may reside not only in code or configuration settings but also in the network architecture itself. That is why the choice of the right topology cannot be considered secondary: the reliability of data storage and transmission directly depends on how the nodes are interconnected. This paper examines the influence of computer network topology on the level of information security. A comparative analysis of four basic topologies — bus, ring, star, and hybrid — was conducted based on such criteria as confidentiality, integrity, availability, fault tolerance, scalability, and intrusion detection complexity. To quantitatively assess vulnerabilities, the authors propose an original model for calculating the information protection coefficient (IPC), grounded in the methodology of the Russian Federal Service for Technical and Export Control (FSTEC). The calculation results show that the hybrid topology provides the highest level of protection (IPC = 0.14), while the star topology (IPC = 0.34) represents the optimal choice for most corporate networks.

Ключевые слова: топология сети, защита информации, коэффициент защиты информации, ФСТЭК, смешанная топология, отказоустойчивость, компьютерные сети.

Keywords: network topology, information security, Information Protection Coefficient, FSTEC, hybrid topology, fault tolerance, computer networks.

Введение

Топология компьютерной сети — это схема соединения устройств: «шина», «звезда», «кольцо» или «смешанная». Она определяет, как данные передаются между узлами, что напрямую влияет на скорость, надёжность и удобство обслуживания. Однако существует ещё один важный параметр, который редко связывают с топологией, — безопасность. В корпоративной среде защиту информации чаще всего строят на криптографии, контроле доступа и мониторинге трафика. Однако на практике уязвимости могут быть заложены уже на уровне физической схемы соединений [1]. Актуальность исследования обусловлена необходимостью количественной оценки влияния архитектурных особенностей сети на её защищённость. Методика анализа защищённости информационных систем (утверждена Федеральной службой по техническому и экспортному контролю 25 ноября 2025 г.) задаёт общие требования к оценке защищённости, однако не рассматривает топологию сети как самостоятельный фактор. В существующих научных работах топологии рассматриваются преимущественно с точки зрения производительности и отказоустойчивости [2], тогда как аспект безопасности остаётся недостаточно изученным [4].

Цель исследования — провести сравнительный анализ топологий компьютерных сетей по показателям безопасности и эксплуатационным характеристикам и разработать количественную модель оценки их защищённости.

Материалы и методы

В работе рассмотрены четыре базовые топологии компьютерных сетей: «шина», «кольцо», «звезда» и «смешанная» (гибридная). Для сравнительного анализа использованы три группы показателей:

Группа 1 — показатели безопасности (табл. 1)

1. Конфиденциальность — возможность несанкционированного перехвата данных.
2. Живучесть при повреждении — способность сохранять работоспособность при отказе отдельных элементов.
3. Защита от сетевых атак типа «отказ в обслуживании» (DDoS).
4. Наличие единой точки отказа.
5. Масштабируемость — возможность расширения сети без существенного снижения характеристик.
6. Сложность обнаружения вторжений.

Группа 2 — эксплуатационные характеристики (табл. 2)

1. Производительность.
2. Задержка передачи данных.
3. Гибкость конфигурации.
4. Помехозащищённость.
5. Сложность обнаружения и локализации неисправностей.

Группа 3 — группы угроз (табл. 3)

1. Перехват трафика.
2. Отказ канала связи.
3. DDoS-атака.
4. Компрометация узла.
5. Единая точка отказа.

Для оценки топологий применяются три методики.

Методика 1 — качественная оценка показателей безопасности и эксплуатационных характеристик (табл. 1 и 2). Каждый показатель оценивается по шкале уровней: «низкий», «средний», «высокий». Оценка отражает степень выраженности соответствующего свойства топологии.

Методика 2 — количественная интерпретация качественных оценок. Уровням присваиваются числовые значения: «низкий» — 1, «средний» — 2, «высокий» — 3. Итоговая оценка топологии по группе показателей определяется как среднее арифметическое значений по соответствующим показателям. Чем выше итоговая оценка, тем лучше характеристика топологии.

Методика 3 — количественная оценка групп угроз по интервальной шкале коэффициента уязвимости от 0 до 1. Каждой угрозе присваивается коэффициент уязвимости:

- 0 — угроза полностью отсутствует;
- 0,1–0,3 — низкая уязвимость;
- 0,4–0,6 — средняя уязвимость;
- 0,7–0,9 — высокая уязвимость.

Значения коэффициентов определены экспертным путём на основе анализа архитектурных особенностей каждой топологии [10, 5]. КЗИ рассчитывается как среднее арифметическое всех коэффициентов уязвимости для каждой топологии. Чем меньше значение КЗИ, тем выше защищённость.

Во всех трёх методиках итоговая оценка топологии определяется как среднее арифметическое значений по соответствующим показателям. Весовые коэффициенты не применяются.

Результаты и обсуждения

Для количественного и наглядного сравнения уязвимостей различных топологий сведём их ключевые характеристики в таблицу 1.

Таблица 1.

Сравнение характеристик топологий (показатели безопасности)

Показатель	Шина	Кольцо	Звезда	Смешанная
Конфиденциальность	Низкая. Данные передаются по общей среде, каждый узел физически может принимать весь трафик сегмента [1].	Средняя. Трафик проходит последовательно через все узлы, что теоретически позволяет перехватывать пакеты при компрометации узла [5].	Высокая. Коммутатор направляет трафик только порту назначения, остальные узлы не видят чужие данные (при отсутствии атак на коммутатор)[3].	Высокая. Используются сегментированные звёзды; трафик внутри сегмента изолирован [6].
Живучесть при повреждении	Низкая. Разрыв общего кабеля в любой точке приводит к потере связи между всеми узлами [2].	Низкая. Разрыв кольца размыкает цепь, сеть распадается на два изолированных сегмента [2].	Высокая. Повреждение кабеля между коммутатором и одним узлом выводит из строя только этот узел [7].	Очень высокая. Отказ одного сегмента не влияет на остальные; критичные сегменты могут резервироваться [6].

Показатель	Шина	Кольцо	Звезда	Смешанная
Защита от DDoS	Низкая. Атака на один узел (например, генерация ложного трафика) может затопить всю шину и парализовать сеть [8].	Средняя. «Зашумление» одного узла нарушает передачу по всему кольцу, но маркерный доступ может частично ограничить влияние [5].	Высокая. Коммутатор может изолировать некорректный порт, локализуя атаку в пределах атакующего узла [7].	Высокая. Локализация атак внутри сегментов; центральные устройства мониторят аномалии [9].
Единая точка отказа	Весь кабель является распределённой точкой отказа — повреждение любого участка фатально [2].	Каждый узел и каждый кабель являются точкой отказа — выход одного узла разрывает кольцо [5].	Центральный коммутатор (или маршрутизатор) — если он выходит из строя, сеть полностью парализована[7].	Коммутаторы могут дублироваться, сохраняя работоспособность [9].
Масштабируемость	Крайне низкая. Добавление новых узлов увеличивает количество потенциальных «слушателей» и снижает надёжность шины [1].	Низкая. Добавление узла требует разрыва кольца, что временно останавливает всю сеть; растёт задержка [4].	Высокая. Новый узел подключается к свободному порту коммутатора без влияния на остальных [10].	Высокая. Расширение происходит путём добавления новых «звёзд» без изменения архитектуры [6].
Сложность обнаружения вторжений	Высокая. Отсутствие центрального управления затрудняет мониторинг; аномальный трафик сложно отличить от нормального [1].	Средняя. Возможен пассивный мониторинг, но локализация источника требует анализа на каждом узле [4].	Низкая. Коммутатор может зеркалировать порты, логировать подключения, быстро выявлять подозрительную активность [7].	Низкая. Централизованные коммутаторы и сегментирование позволяют строить эффективные системы обнаружения вторжений (IDS/IPS) [9].

Количественная интерпретация табл. 1 (Методика 2). Присвоим уровням числовые значения: «низкий» — 1, «средний» — 2, «высокий» — 3. Для показателя «Единая точка отказа» оценка инвертируется: отсутствие единой точки отказа — 3, наличие — 1. Результаты сведены в таблицу 2.

Таблица 2.

Количественная оценка показателей безопасности

Показатель	Шина	Кольцо	Звезда	Смешанная
Конфиденциальность	1	2	3	3
Живучесть при повреждении	1	1	3	3
Защита от DDoS	1	2	3	3
Единая точка отказа (инвертированная)	1	1	1	3
Масштабируемость	1	1	3	3
Сложность обнаружения вторжений (инвертированная)	1	2	3	3
Сумма	**6**	**9**	**16**	**18**
Среднее	**1,00**	**1,50**	**2,67**	**3,00**

Таким образом, по показателям безопасности наилучшей является смешанная топология (средняя оценка 3,00), на втором месте — «звезда» (2,67). Топологии «шина» (1,00) и «кольцо» (1,50) имеют неудовлетворительные показатели.

Таблица 3.

Сравнение дополнительных параметров (эксплуатационные характеристики)

Параметр	Шина	Кольцо	Звезда	Смешанная
Производительность	Низкая. Общая шина — пропускная способность делится между всеми узлами. При росте сети скорость падает катастрофически [2].	Средняя. Данные идут по кругу, каждый узел создаёт микро-задержку. При большом количестве узлов падение скорости заметно [4].	Высокая. Каждый узел имеет выделенный канал до коммутатора. Производительность не падает при добавлении новых узлов (до заполнения портов) [10].	Высокая. Наследует от «звезды» внутри сегментов. При грамотном проектировании падения производительности нет [6].
Задержка передачи данных	Высокая. При коллизиях узлы ждут случайное время и повторяют отправку. Непредсказуемая задержка [2].	Средняя. Пакет проходит последовательно через все узлы, пока не дойдёт до адресата. Задержка растёт с каждым новым узлом [4].	Низкая. Коммутатор получает пакет и мгновенно отправляет только на порт назначения. Задержка минимальна и стабильна [7].	Низкая. В пределах сегмента — как у «звезды». При передаче между сегментами задержка чуть выше, но не критично [6].
Гибкость конфигурации	Очень низкая. Любое изменение требует остановки всей сети и физического вмешательства в общий кабель [1].	Низкая. Добавление или удаление узла требует разрыва кольца и остановки сети. Перенастройка сложная [4].	Высокая. Новый узел подключается к свободному порту коммутатора на ходу. Удаление узла — просто отключить кабель. Сеть продолжает работать [7].	Высокая. Каждый сегмент (отдельная «звезда») настраивается независимо. Можно менять конфигурацию одного сегмента без остановки других [9].
Помехозащищённость	Низкая. Длинный общий кабель работает как антенна. Наводки на одном участке могут исказить сигнал для всех узлов [2].	Средняя. Сигнал регенерируется каждым узлом при передаче дальше, что «очищает» его от помех. Но сам кабель длинный [5].	Высокая. Кабель от коммутатора до узла — отдельная линия. Помеха на одном кабеле не влияет на другие узлы [7].	Высокая. Наследует от «звезды». Каждый сегмент защищён от помех в других сегментах [9].
Сложность обнаружения и локализации неисправностей	Сложная. Обрыв кабеля где-то на шине парализует всю сеть, но понять, в каком именно месте обрыв, трудно. Нужно смотреть каждый участок [1].	Сложная. Сеть разорвана, но неясно, какой именно узел или кабель сломался. Приходится проверять каждый узел и каждый соединительный отрезок [4].	Простая. Коммутатор обычно имеет световую индикацию или интерфейс управления, показывающие, какой порт не работает. Можно быстро заменить кабель или узел [10].	Средняя. При отказе одного сегмента остальные работают. Но локализация требует проверки коммутаторов каждого сегмента по очереди [9].

Количественная интерпретация табл. 3 (Методика 2). Присвоим уровням числовые значения: «низкий» — 1, «средний» — 2, «высокий» — 3. Для параметров «Задержка передачи данных» и «Сложность обнаружения и локализации неисправностей» оценка инвертируется: чем ниже задержка и чем проще обнаружение, тем выше балл. Результаты сведены в таблицу 4.

Таблица 4.

Количественная оценка эксплуатационных характеристик

Параметр	Шина	Кольцо	Звезда	Смешанная
Производительность	1	2	3	3
Задержка передачи данных (инвертированная)	1	2	3	3
Гибкость конфигурации	1	1	3	3
Помехозащищённость	1	2	3	3
Сложность обнаружения и локализации неисправностей (инвертированная)	1	1	3	2
Сумма	**5**	**8**	**15**	**14**
Среднее	**1,00**	**1,60**	**3,00**	**2,80**

Таким образом, по эксплуатационным характеристикам наилучшей является топология «звезда» (средняя оценка 3,00), на втором месте — смешанная топология (2,80). Топологии «шина» (1,00) и «кольцо» (1,60) имеют неудовлетворительные показатели.

Сопоставление результатов по табл. 2 и 4. Смешанная топология лидирует по показателям безопасности (3,00), но немного уступает «звезде» по эксплуатационным характеристикам (2,80 против 3,00). Топология «звезда» обеспечивает наилучшие эксплуатационные характеристики и высокие показатели безопасности (2,67), уступая смешанной только по отказоустойчивости и наличию единой точки отказа. Топологии «шина» и «кольцо» существенно уступают по обеим группам показателей [2, 10].

Расчёт коэффициента защиты информации (КЗИ)

Для количественной оценки защищённости каждой топологии разработана авторская модель на основе интервальной шкалы коэффициента уязвимости от 0 до 1 (Методика 3). В предлагаемой модели каждой топологии ставится в соответствие набор коэффициентов уязвимости по нескольким группам угроз, выделенным на основе анализа архитектурных особенностей [5, 9].

Таблица 5.

Оценка уязвимости топологий по группам угроз

Группа угроз	Шина	Кольцо	Звезда	Смешанная	Обоснование
Перехват трафика	0,9	0,6	0,2	0,1	В «шине» все узлы видят весь трафик; в «кольце» — при компрометации узла; в «звезде» трафик изолирован; в «смешанной» — сегментирован [1, 3].

Группа угроз	Шина	Кольцо	Звезда	Смешанная	Обоснование
Отказ канала связи	0,9	0,8	0,3	0,2	Общий кабель «шины» и «кольца» критичен; в «звезде» отказ только одного узла; в «смешанной» — резервирование [2, 7].
DDoS-атака	0,8	0,5	0,2	0,1	«Шина» парализуется полностью; «кольцо» — частично; «звезда» — коммутатор изолирует порт; «смешанная» — сегментация [8, 9].
Компрометация узла	0,5	0,7	0,1	0,1	В «кольце» каждый узел является ретранслятором; в «шине» — пассивное прослушивание; в «звезде» и «смешанной» — влияние минимально [3, 5].
Единая точка отказа	0,9	0,8	0,9	0,2	«Шина» — весь кабель; «кольцо» — любой узел; «звезда» — центральный коммутатор; «смешанная» — дублирование [2, 7].

Расчёт КЗИ (среднее арифметическое)

1. Шина: КЗИ = (0,9 + 0,9 + 0,8 + 0,5 + 0,9) / 5 = 0,80
2. Кольцо: КЗИ = (0,6 + 0,8 + 0,5 + 0,7 + 0,8) / 5 = 0,68
3. Звезда: КЗИ = (0,2 + 0,3 + 0,2 + 0,1 + 0,9) / 5 = 0,34
4. Смешанная: КЗИ = (0,1 + 0,2 + 0,1 + 0,1 + 0,2) / 5 = 0,14

Сводная таблица результатов

Таблица 6.

Сводные результаты оценки топологий

Топология	Средняя оценка по показателям безопасности (табл. 1а)	Средняя оценка по эксплуатационным характеристикам (табл. 2а)	КЗИ (табл. 3)	Топология
Шина	1,00	1,00	0,80	Шина
Кольцо	1,50	1,60	0,68	Кольцо
Звезда	2,67	3,00	0,34	Звезда

Полученные значения КЗИ отражают интегральную оценку уязвимости каждой топологии. Результаты расчёта показывают, что наилучшую защищённость обеспечивает смешанная топология (КЗИ = 0,14). На втором месте — топология «звезда» (КЗИ = 0,34). Топологии «шина» и «кольцо» имеют неудовлетворительные показатели (0,80 и 0,68 соответственно), что подтверждает их нецелесообразность для построения защищённых сетей [1, 2, 9]. Сопоставление результатов по трём методикам показывает, что смешанная топология обеспечивает наиболее высокий уровень безопасности (средняя оценка 3,00 по табл. 1а, КЗИ = 0,14). Топология «звезда» является оптимальным выбором для большинства корпоративных сетей по соотношению «безопасность / стоимость»: она обеспечивает наилучшие эксплуатационные характеристики (3,00) и высокие показатели безопасности

(2,67) при приемлемом КЗИ (0,34). Топологии «шина» и «кольцо» уступают по всем трём группам показателей и не рекомендуются для использования в современных корпоративных сетях [2, 10, 6].

Заключение

В результате проведённого исследования были получены следующие выводы:

1. Качественный анализ показал, что топологии «шина» и «кольцо» обладают критическими уязвимостями: низкой конфиденциальностью, наличием распределённой единой точки отказа и сложностью обнаружения вторжений. Количественная оценка подтвердила это: средние оценки по показателям безопасности составили 1,00 и 1,50 соответственно, по эксплуатационным характеристикам — 1,00 и 1,60, КЗИ — 0,80 и 0,68. Данные топологии не рекомендуются для использования в современных корпоративных сетях.

2. Топология «звезда» обеспечивает высокую производительность, низкие задержки, гибкость, хорошую помехозащищённость и простоту обнаружения неисправностей. Её единственным существенным недостатком является единая точка отказа — центральный коммутатор, который может быть защищён резервированием. Средняя оценка по показателям безопасности — 2,67, по эксплуатационным характеристикам — 3,00, КЗИ — 0,34.

3. Смешанная топология наследует преимущества «звезды» внутри сегментов и обеспечивает дополнительную отказоустойчивость за счёт сегментации и возможности резервирования критичных компонентов. Средняя оценка по показателям безопасности — 3,00, по эксплуатационным характеристикам — 2,80, КЗИ — 0,14.

4. Количественный расчёт КЗИ с использованием авторской модели на основе интервальной шкалы коэффициента уязвимости подтвердил, что смешанная топология является наиболее защищённой (КЗИ = 0,14), а топология «звезда» (КЗИ = 0,34) — оптимальным выбором для большинства корпоративных сетей по соотношению «безопасность / стоимость».

5. Наиболее высокий уровень безопасности обеспечивает смешанная топология. Полученные результаты могут быть использованы при проектировании защищённых корпоративных сетей и выборе архитектуры, обеспечивающей требуемый уровень безопасности при минимальных затратах на дополнительные средства защиты.

Список литературы

1. Методика анализа защищённости информационных систем: утверждена Федеральной службой по техническому и экспортному контролю 25 ноября 2025 г. – Москва: ФСТЭК России, 2025.
2. Монахова, Э.А., Монахов, О.Г. О некоторых характеристиках циркулянтных и тороидальных структур вычислительных систем // Вестник СибГУТИ. – 2022. – № 4.
3. An Investigation into the Performance of DHKE Algorithm over Different Network Topologies in Cryptography Networks // Proceedings of the. – 2023. – P. 1–6. – DOI: 10.1109/ICISCT60174.2023.10390219.

4. ChameleonNet: Topology Obfuscation Against Tomography With Critical Information Hiding // IEEE Transactions on Networking. – 2025. – T. 33, № 5. – P. 2585–2600. – DOI: 10.1109/TON.2025.11006136.
5. Heidari Safari, M.E., Hosseini, S. A federated framework for simulating and predicting malware propagation in network topologies with privacy awareness // Scientific Reports. – 2026. – DOI: 10.1038/s41598-026-69350-4.
6. Hettiarachchi, E.D. S.I., Sarkar, N.I., & Gutierrez, J. Impact of Southbound Expansion on Clustered OpenFlow Software-Defined Network Controller Synchronisation Using ODL and ONOS // Information. – 2024. – T. 15, № 8. – DOI: 10.3390/info15080440.
7. Li, X., Lee, J., Son, J. & Lee, Y. Tunnel enabled programmable switches obfuscate network topology to defend against link flooding reconnaissance in software defined networking // Scientific Reports. – 2025. – T. 15.
8. Protection Mechanism Against Internal Data Leakage Based on Scitags // Proceedings of the. – 2025. – P. 1–8. – DOI: 10.1109/NAS63823.2025.11181425.
9. Reducing Mean Absolute Error in Attack Predictions With Power-Parametrized Choquet Integrals // ITL – International Journal of Applied Information Technology. – 2026. – DOI: 10.1002/itl2.70368.
10. Thaenchaikun, S., Kanjanasit, K. Evaluation of Secured and Unsecured OSPF Routing Protocols: A Simulated Experiment // East African Journal of Information Technology. – 2026. – T. 9, № 1. – DOI: 10.37284/eajit.9.1.4536.

References

1. [Methodology for analyzing the security of information systems]: approved by the Federal Service for Technical and Export Control on November 25, 2025. Moscow, FSTEC Rossii Publ., 2025. (In Russ.)
2. Monakhova E.A., Monakhov O.G. [On some characteristics of circulant and toroidal structures of computing systems]. Vestnik SibGUTI, 2022, no. 4. (In Russ.)
3. An Investigation into the Performance of DHKE Algorithm over Different Network Topologies in Cryptography Networks // Proceedings of the. – 2023. – P. 1–6. – DOI: 10.1109/ICISCT60174.2023.10390219.
4. ChameleonNet: Topology Obfuscation Against Tomography With Critical Information Hiding. IEEE Transactions on Networking, 2025, vol. 33, no. 5, P. 2585–2600. DOI: 10.1109/TON.2025.11006136.
5. Heidari Safari M.E., Hosseini S. A federated framework for simulating and predicting malware propagation in network topologies with privacy awareness. Scientific Reports. – 2026. – DOI: 10.1038/s41598-026-69350-4.
6. Hettiarachchi E.D.S.I., Sarkar N.I., Gutierrez J. Impact of Southbound Expansion on Clustered OpenFlow Software-Defined Network Controller Synchronisation Using ODL and ONOS. Information, 2024, vol. 15, no. 8, P. 440. DOI: 10.3390/info15080440.

7. Li X., Lee J., Son J., Lee Y. Tunnel enabled programmable switches obfuscate network topology to defend against link flooding reconnaissance in software defined networking. Scientific Reports, 2025, vol. 15.
8. Protection Mechanism Against Internal Data Leakage Based on Scitags // Proceedings of the. – 2025. – P. 1–8. – DOI: 10.1109/NAS63823.2025.11181425.
9. Reducing Mean Absolute Error in Attack Predictions With Power-Parametrized Choquet Integrals. ITL – International Journal of Applied Information Technology. – 2026. – DOI: 10.1002/itl2.70368.
10. Thaenchaikun S., Kanjanasit K. Evaluation of Secured and Unsecured OSPF Routing Protocols: A Simulated Experiment. East African Journal of Information Technology, 2026, vol. 9, no. 1. DOI: 10.37284/eajit.9.1.4536.

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

Статья поступила в редакцию: 24.08.2026
Статья принята к публикации: 14.09.2026
Статья опубликована: 28.09.2026

УДК 004.8+004.9

DOI: 10.32743/UniTech.2026.150.9.23412

Моделирование процесса формирования визуального контента в интеллектуальных системах искусственного интеллекта на основе Prompt Engineering

Ганиев Акмал Абдухалилович
ст. преп. кафедры управления промышленностью и цифровых технологий
Международный университет Нордик
Республика Узбекистан, г. Ташкент
ORCID: 0009-0006-9402-1508
E-mail: g.akmal@nordicuniversity.org

Modeling the Process of Visual Content Generation in Intelligent AI Systems Based on Prompt Engineering

Ganiev Akmal Abdukhalilovich
senior Lecturer
Department of Industrial Management and Digital Technologies Nordic International University
Republic of Uzbekistan, Tashkent

Аннотация

В статье рассматривается процесс формирования визуального контента в интеллектуальных системах искусственного интеллекта на основе технологий Prompt Engineering. Актуальность исследования обусловлена необходимостью повышения точности преобразования естественно-языковых пользовательских запросов в визуальный контент с учетом заданных характеристик, контекста и требований к изображению. Проведен анализ существующих подходов к генерации изображений с использованием больших языковых моделей и современных генеративных нейронных сетей. Предложена авторская концептуальная модель интеллектуального формирования визуального контента, предусматривающая последовательные этапы получения исходного запроса, его семантического анализа, структуризации, оптимизации и последующей оценки сформированного изображения. В качестве структурной основы промпта выделены четыре

Библиографическое описание: Ганиев А.А. Моделирование процесса формирования визуального контента в интеллектуальных системах искусственного интеллекта на основе Prompt Engineering // Universum: технические науки : электрон. научн. журн. 2026. 9(150). URL: https://7universum.com/ru/tech/archive/item/23412

взаимосвязанных компонента: содержание визуальной сцены (S), визуальные характеристики (V), пользовательские требования (U) и контекст формирования (C). Разработан алгоритм интеллектуальной оптимизации промпта, предусматривающий проверку полноты и согласованности его компонентов, уточнение параметров и итерационную корректировку запроса на основе полученного результата. Для оценки эффективности предложенного подхода использованы критерии семантического соответствия, полноты визуального описания, стилевой согласованности, детализации и устойчивости результатов при повторной генерации. Экспериментальная проверка проведена с использованием генеративных систем ChatGPT Images 2.0, Pixmind, Qwen 3.0 Pro и FLUX.2 [klein]. Полученные результаты подтверждают положительное влияние структуризации и оптимизации промптов на качество формируемого визуального контента. Предлагаемый подход может быть использован при разработке интеллектуальных систем поддержки проектирования, цифрового дизайна, образовательных платформ и мультимедийных приложений.

Abstract

The article examines the process of visual content generation in intelligent artificial intelligence systems based on Prompt Engineering technologies. The relevance of the study is determined by the need to improve the accuracy of transforming natural-language user queries into visual content while taking into account specified characteristics, contextual information, and user requirements. Existing approaches to image generation using large language models and modern generative neural networks are analyzed. An original conceptual model for intelligent visual content generation is proposed, incorporating sequential stages of obtaining the initial query, performing semantic analysis, structuring and optimizing the prompt, and subsequently evaluating the generated image. Four interconnected components are identified as the structural basis of the prompt: visual scene content (S), visual characteristics (V), user requirements (U), and generation context (C). An algorithm for intelligent prompt optimization is developed, including verification of the completeness and consistency of its components, parameter refinement, and iterative query correction based on the generated result. The effectiveness of the proposed approach is evaluated using the criteria of semantic correspondence, completeness of visual description, stylistic consistency, image detail, and stability of results during repeated generation. Experimental verification was conducted using ChatGPT Images 2.0, Pixmind, Qwen 3.0 Pro, and FLUX.2 [klein]. The obtained results confirm the positive effect of prompt structuring and optimization on the quality of generated visual content. The proposed approach can be applied in the development of intelligent design support systems, digital design tools, educational platforms, and multimedia applications.

Ключевые слова: искусственный интеллект, Prompt Engineering, генеративный искусственный интеллект, визуальный контент, генерация изображений, диффузионные модели, интеллектуальные системы, моделирование.

Keywords: Artificial Intelligence, Prompt Engineering, Generative AI, Visual Content, Image Generation, Diffusion Models, Intelligent Systems, Modeling.

Введение

В последние годы искусственный интеллект (ИИ) стал одной из ключевых технологий цифровой трансформации, оказывая существенное влияние на обработку информации, автоматизацию интеллектуальной деятельности и создание цифрового контента [7, 9]. Особое развитие получили генеративные модели искусственного интеллекта, основанные на использовании больших языковых моделей и диффузионных нейронных сетей, позволяющих автоматически создавать текстовые и графические материалы по естественно-языковым запросам пользователя [2, 6, 10]. Современные системы, такие как ChatGPT Images 2.0, Pixmind, Qwen 3.0 Pro, FLUX.2 [klein] активно применяются в цифровом дизайне, промышленном проектировании, образовании и мультимедийных технологиях [3].

Одним из ключевых факторов, определяющих качество работы генеративных моделей, является технология Prompt Engineering, обеспечивающая структурированное формирование текстовых запросов для получения требуемого результата [4]. Качество создаваемого визуального контента во многом определяется корректностью построения промпта, полнотой контекстной информации и точностью описания объекта генерации [5].

Несмотря на значительное количество исследований в области генеративного искусственного интеллекта, основное внимание уделяется совершенствованию архитектур нейронных сетей и методов генерации изображений [2, 3, 6, 10, 8]. При этом вопросы моделирования процесса преобразования текстового промпта в визуальный контент, включая этапы семантического анализа, структурной обработки запроса и оценки качества результатов генерации, остаются недостаточно изученными. Отсутствие комплексной модели взаимодействия пользователя с интеллектуальной системой определяет актуальность настоящего исследования.

Ранее автором были рассмотрены отдельные аспекты применения технологий искусственного интеллекта и методов Prompt Engineering при разработке цифрового образовательного контента и интеллектуальных информационных решений [1]. Настоящая работа является продолжением этих исследований и направлена на разработку модели формирования визуального контента в интеллектуальных системах искусственного интеллекта.

Целью исследования является разработка модели процесса формирования визуального контента в интеллектуальных системах искусственного интеллекта на основе технологии Prompt Engineering, обеспечивающей повышение качества интерпретации текстовых запросов и эффективности генерации изображений.

Задачи

1. провести анализ современных методов формирования визуального контента в интеллектуальных системах искусственного интеллекта, исследовать особенности генеративных моделей и определить роль технологии Prompt Engineering в процессе обработки текстовых запросов;

2. исследовать существующие подходы к структурированию текстовых промптов, выявить факторы, влияющие на точность интерпретации пользовательских запросов и качество генерируемых изображений;

3. разработать концептуальную модель процесса формирования визуального контента на основе технологии Prompt Engineering, включающую этапы семантического анализа, структурной обработки промпта, генерации изображения и оценки качества результата;

4. предложить алгоритм интеллектуальной обработки текстовых промптов, обеспечивающий повышение эффективности взаимодействия пользователя с генеративной системой искусственного интеллекта;

5. разработать математическую модель оценки качества формируемого визуального контента с учетом семантического соответствия, полноты описания, стилевой согласованности, детализации изображения и устойчивости результата при повторной генерации;

6. провести экспериментальное исследование предложенной модели с использованием современных генеративных систем искусственного интеллекта и выполнить сравнительный анализ полученных результатов по показателям качества визуального контента.

Материалы и методы

Объектом исследования является процесс формирования визуального контента в интеллектуальных системах искусственного интеллекта на основе технологии Prompt Engineering. Предмет исследования включает методы структурирования и семантической обработки текстовых промптов, а также механизмы преобразования естественно-языковых запросов в изображения с использованием современных генеративных моделей.

Материалами исследования послужили научные публикации по генеративному искусственному интеллекту, Prompt Engineering, большим языковым и диффузионным моделям, а также результаты экспериментального применения систем ChatGPT Images 2.0, Pixmind, Qwen 3.0 Pro и FLUX.2 [klein].

Методологическую основу составили системный и сравнительный анализ, концептуальное моделирование, семантический анализ и структурное проектирование. Предлагаемая модель предусматривает последовательную обработку запроса: семантический анализ, оптимизацию промпта, генерацию изображения и оценку результата.

Эффективность модели оценивалась по пяти критериям: семантическому соответствию, полноте визуального описания, стилевой согласованности, детализации и устойчивости результата при повторной генерации. Сравнение проводилось при сопоставимых условиях и одинаковых исходных заданиях до и после оптимизации промптов.

Таблица 1.

Методы исследования

Метод	Назначение
Системный анализ	Исследование структуры процесса формирования визуального контента
Семантический анализ	Интерпретация текстовых промптов
Концептуальное моделирование	Разработка авторской модели процесса
Сравнительный анализ	Сопоставление генеративных AI-систем
Экспериментальный метод	Оценка качества результатов генерации

Результаты и обсуждение

2.1. Концептуальная модель процесса формирования визуального контента

Результаты исследования позволили сформировать концептуальную модель преобразования естественно-языкового пользовательского запроса в визуальный контент. В отличие от прямой схемы «текстовый запрос – изображение», предлагаемая модель предусматривает последовательные этапы семантической и структурной обработки, оптимизации и оценки результата.

Анализ текстовых промптов показал, что их структура и полнота существенно влияют на качество генерируемого визуального контента. Недостаточно структурированный запрос может приводить к неполному представлению объектов, нарушению их взаимного расположения и несоответствию заданному стилю. С учетом выявленных особенностей процесс формирования визуального контента представлен следующей последовательностью:

Рисунок 1. Процесс генерации визуального контента

На первом этапе формируется исходный пользовательский запрос, после чего выполняется его семантический анализ с выделением основных объектов, атрибутов, контекста, пространственных отношений и визуальных ограничений.

На этапе структуризации выделенная информация преобразуется в упорядоченную структуру промпта, после чего выполняется его оптимизация путем уточнения параметров, устранения неоднозначностей и дополнения необходимой контекстной информации. Оптимизированный промпт передается генеративной модели для формирования визуального контента.

Полученный результат оценивается по критериям семантического соответствия, полноты, стилевой согласованности, детализации и устойчивости при повторной генерации. При выявлении несоответствий выполняется корректировка промпта и повторная генерация. Таким образом, предложенная модель имеет итерационный характер:

$$P_i \rightarrow G(P_i) \rightarrow K \rightarrow P_{i+1}$$

где P_i — текущий промпт, $G(P_i)$ — результат его обработки генеративной моделью, K_i — оценка качества полученного визуального результата, а P_{i+1} — скорректированный промпт, сформированный с учетом выявленных несоответствий.

Важным элементом предложенной модели является обратная связь между результатом генерации и последующей обработкой текстового запроса. Она обеспечивает последовательное уточнение промпта с учетом выявленных несоответствий и повышение соответствия визуального результата исходному пользовательскому замыслу.

Таким образом, качество визуального контента определяется характеристиками исходного запроса, структурой и качеством семантической обработки промпта, а также возможностями генеративной модели. Prompt Engineering в предложенной модели выступает промежуточным интеллектуальным механизмом, связывающим пользовательское намерение с процессом визуального синтеза.

2.2. Структурная модель Prompt Engineering

Для повышения точности формирования визуального контента промпт рассматривается как структурированный набор взаимосвязанных компонентов, отражающих содержание изображения, его визуальные характеристики, требования пользователя и контекст. Такой подход позволяет перейти от свободного текстового описания к формализованной структуре. В рамках исследования структура промпта представляется следующим образом:

$$P=\{S,V,U,C\}$$

где (P) — структурированный промпт; (S) — содержание или основной объект визуальной сцены; (V) — визуальные характеристики; (U) — пользовательские требования и ограничения; (C) — контекст формирования визуального контента.

Компонент (S) (Subject) определяет содержание изображения: основные объекты, их действия и взаимное расположение.

Компонент (V) (Visual characteristics) характеризует визуальное представление: художественный стиль, цветовую гамму, освещение, композицию, перспективу и детализацию.

Компонент (U) (User requirements) отражает требования и ограничения пользователя, включая формат, соотношение сторон, степень реалистичности и наличие заданных элементов.

Компонент (C) (Context) определяет контекст формирования изображения: место и время действия, назначение контента и окружающую среду.

Таким образом, структурированный промпт может рассматриваться как совокупность четырех взаимосвязанных компонентов:

$$P=\{S,V,U,C\}$$

Компоненты S, V, U и C взаимосвязаны: изменение одного компонента может потребовать корректировки других. Поэтому качество изображения определяется не только полнотой отдельных компонентов, но и согласованностью их содержания.

Предлагаемая структура создает основу для интеллектуальной оптимизации промпта, позволяя выявлять отсутствующие или противоречивые параметры и формировать уточненный запрос перед его передачей генеративной модели.

Таким образом, модель $P=\{S,V,U,C\}$ выступает промежуточным уровнем между естественно-языковым запросом и генеративной моделью, обеспечивая структурированное представление пользовательского намерения и автоматизированную оптимизацию промпта.

2.3. Алгоритм интеллектуальной оптимизации промпта

На основе разработанной структурной модели предложен алгоритм интеллектуальной оптимизации текстового промпта, направленный на повышение точности интерпретации пользовательского запроса и качества визуального контента. Алгоритм предусматривает последовательную обработку содержательных, визуальных, пользовательских и контекстных характеристик запроса. Алгоритм включает следующие основные этапы:

1. Получение исходного запроса — прием естественно-языкового описания требуемого визуального контента.

2. Семантический анализ — определение объектов, атрибутов, действий, взаимосвязей и контекста, а также выявление неоднозначностей.

3. Декомпозиция запроса — распределение выделенной информации по компонентам S, V, U и C.

4. Проверка полноты и согласованности — выявление отсутствующих параметров и логических противоречий между компонентами.

5. Дополнение и уточнение параметров — устранение выявленных неопределенностей с сохранением исходного пользовательского намерения.

6. Формирование оптимизированного промпта — объединение обработанных компонентов в единый структурированный запрос.

7. Генерация и оценка результата — передача промпта генеративной модели и оценка полученного изображения по установленным критериям качества.

8. Корректировка промпта — при выявлении несоответствий корректируются соответствующие компоненты и выполняется повторная генерация. Обобщенно алгоритм может быть представлен следующей последовательностью:

$$Q \rightarrow O\{S,V,U,C\} \rightarrow P* \rightarrow G(P*) \rightarrow K$$

где Q — исходный пользовательский запрос; S,V,U,C — компоненты структурного представления запроса; P∗ — оптимизированный промпт; G(P∗) — результат генерации; K — оценка качества визуального контента.

В отличие от непосредственной передачи исходного запроса генеративной модели, предложенный алгоритм предусматривает предварительную проверку и структурную обработку информации, что позволяет снизить влияние неоднозначных формулировок и повысить полноту промпта. Проверка согласованности компонентов позволяет выявлять потенциальные противоречия и отсутствующие параметры до начала генерации.

Таким образом, алгоритм обеспечивает переход от свободного естественно-языкового описания к структурированному и оптимизированному промпту и формирует итерационный процесс, в котором результаты генерации используются для последующей корректировки запроса. Разработанный алгоритм служит основой для экспериментальной проверки предложенного подхода и сравнительного анализа результатов генерации исходных и оптимизированных промптов.

2.4. Математическая модель оценки качества визуального контента

Для количественной оценки эффективности предложенного подхода определены критерии соответствия визуального контента исходному пользовательскому запросу. Качество результата оценивается по пяти взаимосвязанным критериям: семантическому соответствию, полноте визуального описания, стилевой согласованности, детализации и устойчивости результата при повторной генерации. На их основе интегральный показатель качества определяется следующим образом:

$$K=w_1K_s+w_2K_c+w_3K_{st}+w_4K_d+w_5K_r$$

где K — интегральная итоговая оценка качества визуального контента; K_s — показатель семантического соответствия изображения исходному промпту; K_c — показатель полноты представления заданных элементов; K_{st} — показатель стилевой согласованности; K_d — показатель детализации изображения; K_r — показатель устойчивости результата при повторной генерации; $w_1,w_2,\ldots,w_5$ — весовые коэффициенты, определяющие вклад соответствующих критериев в итоговую оценку качества.

Для обеспечения сопоставимости критериев оценки значения показателей K_s, K_c, K_{st}, K_d и K_r принимаются в диапазоне от 0 до 1. Значение, близкое к 0, соответствует низкой степени выполнения соответствующего критерия, тогда как значение, близкое к 1, характеризует высокий уровень соответствия установленным требованиям.

Весовые коэффициенты w_1, w_2, w_3, w_4, w_5 определяют относительную значимость каждого критерия при формировании итоговой оценки качества. В рамках настоящего исследования для всех пяти критериев установлены одинаковые весовые коэффициенты, что

позволяет обеспечить сопоставимость результатов различных генеративных систем и исключить дополнительное влияние субъективного определения приоритетности отдельных критериев. Поэтому принимается $w_1 = w_2 = w_3 = w_4 = w_5 = 0{,}2$.

Таким образом, интегральный показатель K объединяет характеристики визуального результата в единую оценку качества, при этом более высокое значение K свидетельствует о большем соответствии изображения исходному пользовательскому запросу.

Для проверки эффективности предложенного подхода далее проводится сравнительный анализ результатов генерации исходных и оптимизированных промптов при одинаковых условиях и установленных критериях оценки.

2.5. Экспериментальное исследование и сравнительный анализ

Для экспериментальной проверки предложенной модели проведено сравнительное исследование с использованием систем ChatGPT Images 2.0, Pixmind, Qwen 3.0 Pro и FLUX.2 [klein]. Эксперимент выполнялся при одинаковых условиях и идентичных исходных заданиях. Оптимизированные промпты формировались в соответствии со структурной моделью P={S,V,U,C}.

В качестве базового тестового задания использовался следующий пользовательский запрос:

«Создать фотореалистичное изображение современной университетской аудитории, в которой преподаватель проводит занятие по искусственному интеллекту для студентов. На экране отображается схема нейронной сети, аудитория имеет современный технологичный интерьер, естественное освещение, реалистичные цвета, горизонтальная композиция».

Для визуального сопоставления на рисунке 2 представлены три повторные генерации по исходному и три по оптимизированному промпту, полученные с использованием FLUX.2 [klein]. Сравнение позволяет оценить изменения детализации, полноты представления объектов, стилевой согласованности и устойчивости генерации после оптимизации промпта

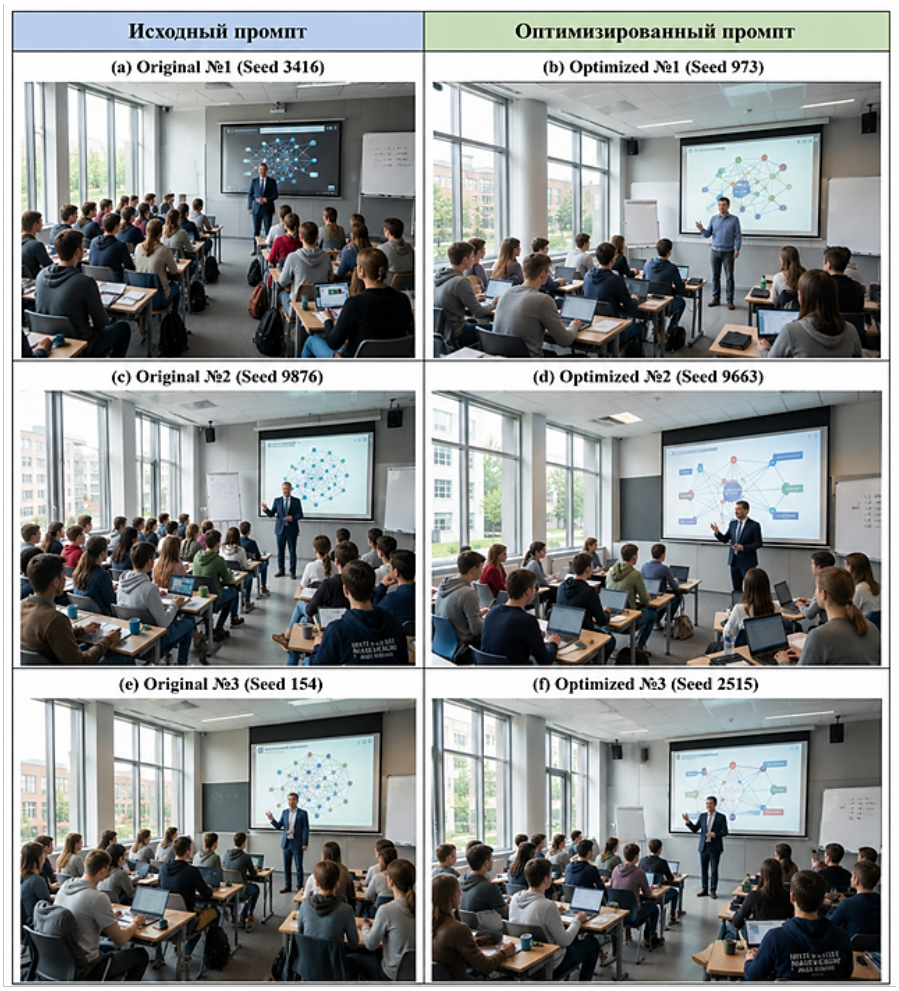

Рисунок 2. Сравнение результатов генерации FLUX.2 [klein] по исходному и оптимизированному промптам

Для дальнейшей обработки исходный пользовательский запрос был декомпозирован в соответствии с четырьмя компонентами структурной модели $P=\{S,V,U,C\}$. Результаты структуризации представлены в таблице 2.

Таблица 2.

Структуризация исходного пользовательского запроса

Компонент	Содержание
S – Содержание	Современная университетская аудитория, преподаватель, студенты, проведение занятия по искусственному интеллекту, схема нейронной сети на экране
V – Визуальные характеристики	Фотореалистичный стиль, современный технологичный интерьер, естественное освещение, реалистичные цвета, горизонтальная композиция
U – Пользовательские требования	Фотореалистичное представление, четкое отображение основных объектов, горизонтальный формат изображения
C – Контекст	Образовательная среда университета, учебное занятие по искусственному интеллекту, использование цифровых технологий

На основе проведенной структуризации был сформирован оптимизированный вариант промпта:

«Фотореалистичное изображение современной университетской аудитории, в которой преподаватель проводит занятие по искусственному интеллекту для студентов. На большом экране отображается схема нейронной сети с четко различимыми узлами и связями. Студенты находятся за учебными столами с ноутбуками и внимательно слушают преподавателя. Аудитория имеет современный технологичный интерьер, естественное освещение, реалистичную цветовую гамму, высокую детализацию и горизонтальную композицию. Широкий угол обзора, четкое расположение основных объектов, реалистичная перспектива».

Для оценки результатов использовались пять критериев: семантическое соответствие, полнота визуального описания, стилевая согласованность, детализация и устойчивость при повторной генерации. Для единообразия оценки применялась пятиуровневая шкала, представленная в таблице 3. Оценка проводилась автором исследования на основе визуального сопоставления результатов генерации с исходным заданием.

Таблица 3.

Шкала оценки критериев качества визуального контента

Значение показателя	Уровень соответствия
0,0–0,2	Очень низкий
>0,2–0,4	Низкий
>0,4–0,6	Средний
>0,6–0,8	Высокий
>0,8–1,0	Очень высокий

В рамках исследования всем пяти критериям присвоены одинаковые весовые коэффициенты $w_1=w_2=w_3=w_4=w_5=0{,}2$, что обеспечивает сопоставимость результатов различных генеративных систем. При этом сумма весовых коэффициентов составляет 1, а интегральный показатель K определяется как среднее значение пяти нормированных критериев:

$$K = 0,2K_s + 0,2K_c + 0,2K_{st} + 0,2K_d + 0,2K_r$$

Эксперимент проводился в два этапа: сначала генеративным системам передавался исходный промпт, затем он подвергался семантическому анализу и структуризации по компонентам S,V,U,C, после чего формировался оптимизированный промпт и выполнялась повторная генерация.

Рисунок 3. Сравнение результатов генерации исходного и оптимизированного промптов

Для визуального сопоставления результатов ChatGPT Images 2.0 на рисунке 3 представлены три генерации по исходному и три по оптимизированному промпту. Сравнение позволяет оценить изменения детализации, представления ключевых объектов и соответствия заданным характеристикам

Как видно на рисунке 3, использование оптимизированного промпта в ChatGPT Images 2.0 обеспечивает более полное представление основных элементов сцены — преподавателя, студентов, экрана с визуализацией нейронной сети и современной учебной среды, а также повышает детализацию визуального результата.

Для обеспечения сопоставимости во всех системах использовались одинаковые исходные задания и структура требований. При сравнении учитывались визуальные характеристики и степень сохранения исходного пользовательского замысла после оптимизации промпта. Для дополнительной проверки предложенного подхода аналогичный эксперимент проведен в системе Pixmind

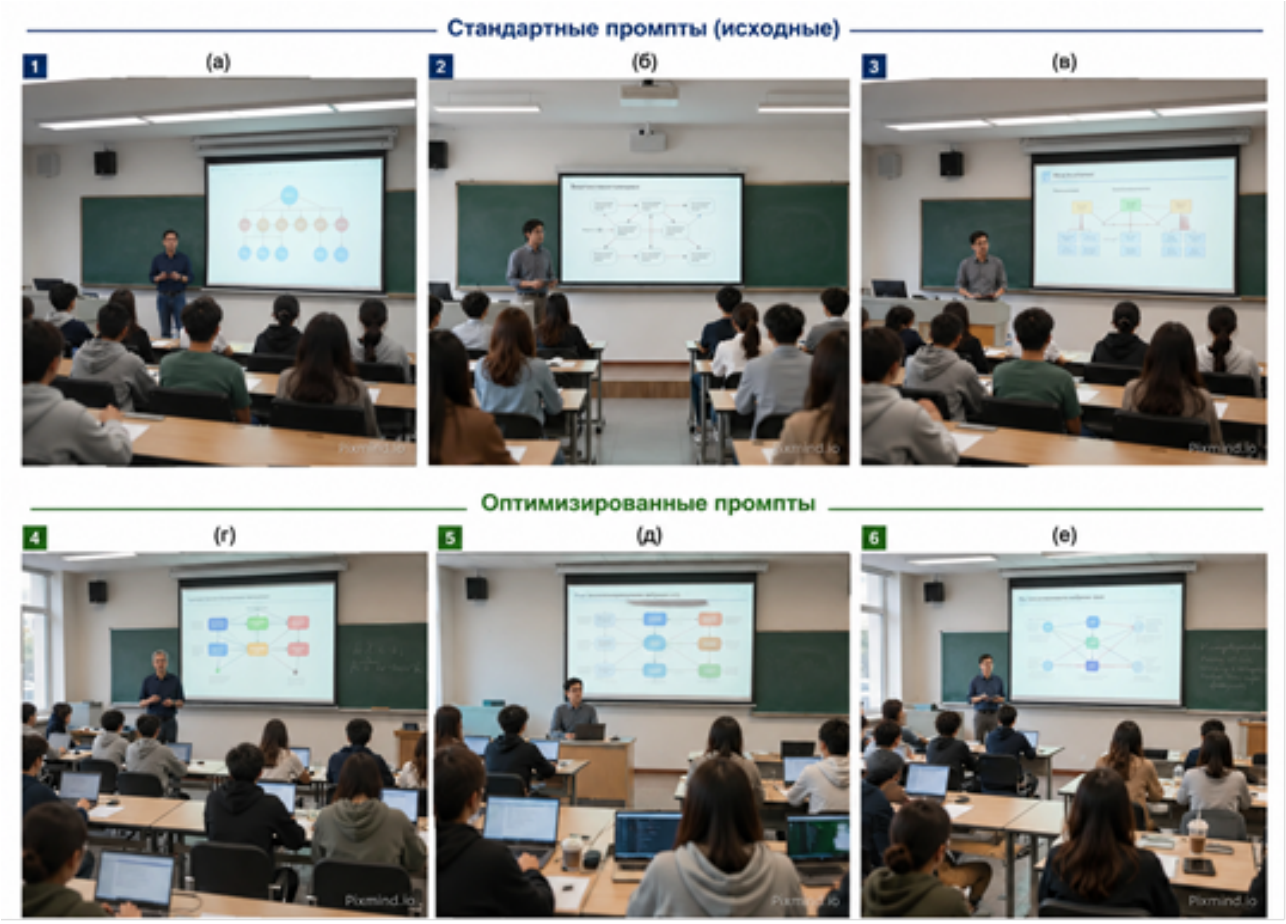

Рисунок 4. Сравнение результатов генерации Pixtmind по исходному и оптимизированному промптам

На рисунке 4 представлены три генерации по исходному и три по оптимизированному промпту. Сравнение позволяет оценить влияние оптимизации на полноту представления объектов, композиционную согласованность и соответствие заданному визуальному контексту.

Для дополнительной визуальной проверки предложенного подхода аналогичный эксперимент был проведен с использованием системы Qwen 3.0 Pro. На рисунке 5 представлены три повторные генерации по исходному промпту и три повторные генерации по оптимизированному промпту. Сопоставление результатов позволяет оценить влияние структуризации и оптимизации промпта на полноту представления объектов, композиционную согласованность, детализацию и соответствие заданному визуальному контексту.

Рисунок 5. Сравнение результатов генерации Qwen 3.0 Pro по исходному и оптимизированному промптам

В рамках сравнительного анализа сопоставлялись результаты, полученные при использовании исходного и семантически и структурно оптимизированного промптов. Такой подход позволяет оценить влияние предложенной модели P={S,V,U,C} на качество формируемого визуального контента [1].

Результаты оценивались по пяти критериям: семантическому соответствию, полноте визуального описания, стилевой согласованности, детализации и устойчивости при повторной генерации. На их основе формировалась интегральная оценка качества, представленная в таблице 4.

Таблица 4.

Сравнительная оценка результатов генерации исходных и оптимизированных промптов

Генеративная система	Вариант промпта	K_s	K_c	K_{st}	K_d	K_r	K
ChatGPT Images 2.0	Исходный	0,7	0,6	0,8	0,7	0,8	0,72
ChatGPT Images 2.0	Оптимизированный	0,9	0,9	0,9	0,9	0,9	0,90
Pixmind	Исходный	0,84	0,83	0,86	0,79	0,85	0,83
Pixmind	Оптимизированный	0,88	0,87	0,89	0,83	0,90	0,87
Qwen 3.0 Pro	Исходный	0,85	0,84	0,82	0,85	0,80	0,83
Qwen 3.0 Pro	Оптимизированный	0,94	0,95	0,94	0,96	0,90	0,94
FLUX.2 [klein]	Исходный	0,90	0,85	0,90	0,90	0,90	0,89
FLUX.2 [klein]	Оптимизированный	0,95	0,95	0,90	0,90	0,90	0,92

Результаты экспериментального исследования показывают, что применение оптимизированных промптов привело к повышению интегрального показателя качества K во всех исследуемых генеративных системах. Для ChatGPT Images 2.0 показатель K увеличился с 0,72 до 0,90, что соответствует приросту на 25,0 %. Для Pixmind показатель увеличился с 0,83 до 0,87, или на 4,8 %. В системе Qwen 3.0 Pro показатель K повысился с 0,83 до 0,94, что соответствует приросту на 13,3 %. Для FLUX.2 [klein] показатель увеличился с 0,89 до 0,92, или на 3,4 %. Наибольший прирост интегрального показателя наблюдается в системе ChatGPT Images 2.0 (+25,0 %), тогда как наименьший прирост зафиксирован в системе FLUX. 2 [klein] (+3,4 %). В целом полученные результаты подтверждают положительное влияние структуризации и оптимизации промптов на качество формируемого визуального контента.

Заключение

В результате исследования разработана модель формирования визуального контента на основе технологии Prompt Engineering, включающая семантический анализ, структуризацию, оптимизацию промпта, генерацию и оценку результата. Предложена структурная модель промпта $P=\{S,V,U,C\}$ и интегральный показатель качества K, основанный на пяти критериях оценки.

Экспериментальная проверка с использованием ChatGPT Images 2.0, Pixmind, Qwen 3.0 Pro и FLUX.2 [klein] показала повышение показателя K при применении оптимизированных промптов: соответственно на 25,0 %, 4,8 %, 13,3 % и 3,4 %. Полученные результаты свидетельствуют о положительном влиянии структуризации и оптимизации промптов на качество формируемого визуального контента. Предложенная модель может применяться в интеллектуальных системах проектирования, цифрового дизайна, образования и мультимедийных приложениях.

Список литературы

1. Ганиев А.А. Моделирование визуализации на основе промптов в системах искусственного интеллекта // Научные основы и современные проблемы интеграции медиа и искусственного интеллекта: сборник докладов Международной научно-практической конференции. – 2026.
2. Cao P., Zhou F., Song Q., Yang L. Controllable Generation With Text-to-Image Diffusion Models: A Survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2026. – Т. 48, № 4. – P. 4771–4791.
3. Chen Z., Zhang L., Weng F., Pan L., Lan Z. Tailored Visions: Enhancing Text-to-Image Generation with Personalized Prompt Rewriting // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024.
4. Gu J., Han Z., Chen S., Beirami A., He B., Zhang G., Liao R., Qin Y., Tresp V., Torr P.H.S. A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models. – 2023. – 15 P.

5. Mahajan S., Rahman T., Yi K.M., Sigal L. Prompting Hard or Hardly Prompting: Prompt Inversion for Text-to-Image Diffusion Models // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024.
6. Mo W., Zhang T., Bai Y., Su B., Wen J.-R., Yang Q. Dynamic Prompt Optimizing for Text-to-Image Generation // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024.
7. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2022.
8. Yao J. et al. PromptCoT: Align Prompt Distribution via Adapted Chain-of-Thought // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024.
9. Zhang C., Zhang C., Zhang M., Kweon I.S., Kim J. Text-to-image Diffusion Models in Generative AI: A Survey. – 2023. – 14 P.
10. Zhang L., Rao A., Agrawala M. Adding Conditional Control to Text-to-Image Diffusion Models // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2023.

References

1. Ganiev A.A. Modeling Prompt-Based Visualization in Artificial Intelligence Systems // Nauchnye osnovy i sovremennye problemy integratsii media i iskusstvennogo intellekta: sbornik dokladov Mezhdunarodnoj nauchno-prakticheskoj konferentsii. – 2026. – P. 95–98. (In Russ.)
2. Cao P., Zhou F., Song Q., Yang L. Controllable Generation With Text-to-Image Diffusion Models: A Survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2026. – Vol. 48, No. 4. – P. 4771–4791.
3. Chen Z., Zhang L., Weng F., Pan L., Lan Z. Tailored Visions: Enhancing Text-to-Image Generation with Personalized Prompt Rewriting // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024.
4. Cao P., Zhou F., Song Q., Yang L. Controllable Generation With Text-to-Image Diffusion Models: A Survey // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2026. – Vol. 48, No. 4. – P. 4771–4791.
5. Mahajan S., Rahman T., Yi K.M., Sigal L. Prompting Hard or Hardly Prompting: Prompt Inversion for Text-to-Image Diffusion Models // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024.
6. Mo W., Zhang T., Bai Y., Su B., Wen J.-R., Yang Q. Dynamic Prompt Optimizing for Text-to-Image Generation // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024.

7. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2022.
8. Yao J. et al. PromptCoT: Align Prompt Distribution via Adapted Chain-of-Thought // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024.
9. Zhang C., Zhang C., Zhang M., Kweon I.S., Kim J. Text-to-image Diffusion Models in Generative AI: A Survey. – 2023. – 14 P.
10. Zhang L., Rao A., Agrawala M. Adding Conditional Control to Text-to-Image Diffusion Models // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2023.

Статья поступила в редакцию: 12.09.2026
Статья принята к публикации: 18.09.2026
Статья опубликована: 28.09.2026

УДК 004.89+004.4

DOI: 10.32743/UniTech.2026.150.9.23398

Сравнение эмбеддинг-моделей для семантического кэширования запросов к большим языковым моделям

Колышкин Виктор Викторович
инженер-программист
независимый исследователь
Российская Федерация, г. Москва
ORCID: 0009-0005-2632-5107
E-mail: mr.viktor.kolyshkin@gmail.com

A comparison of embedding models for semantic caching of large language model requests

Kolyshkin Viktor
Software engineer, independent researcher
Russia, Moscow

Аннотация

Обращения к большим языковым моделям (LLM) через API дороги и медленны, а часть запросов повторяется по смыслу; семантический кэш возвращает готовый ответ на близкий запрос, снижая стоимость и задержку. Ключевой выбор — эмбеддинг-модель и порог сходства — управляет скрытым компромиссом между долей попаданий и долей ложных попаданий (кэш вернул ответ на другой запрос). В работе на публичном наборе Quora Question Pairs (40 430 пар) сравниваются три компактные модели (all-MiniLM-L6-v2, e5-small-v2, bge-small-en-v1.5) и два лексических базиса (точное совпадение, TF-IDF) по совместным осям: ROC/AUC, доля попаданий при заданном бюджете ложных попаданий, латентность и пропускная способность на CPU и GPU, а также оценка денежной экономии. Семантические модели (AUC 0,86–0,88) заметно превосходят лексические (TF-IDF 0,73; точное совпадение ловит менее 0,5 % смысловых дублей). Модель bge-small даёт лучшее качество, но уступает MiniLM по скорости на CPU; при строгом бюджете ложных попаданий (1 %) в попарной оценке отделяется лишь около одной шестой повторов. Стоимость эмбеддинга пренебрежимо мала относительно экономии на вызовах LLM. Отдельно

Библиографическое описание: Колышкин В.В. Сравнение эмбеддинг-моделей для семантического кэширования запросов к большим языковым моделям // Universum: технические науки : электрон. научн. журн. 2026. 9(150). URL: https://7universum.com/ru/tech/archive/item/23398

показано, что операционная доля ложных попаданий растёт с размером кэша и превышает попарную оценку. Результаты дают инженеру основу для выбора модели и порога под целевой профиль.

Abstract

Calls to large language models (LLMs) via API are costly and slow, while a share of requests repeats in meaning; a semantic cache returns a stored answer for a similar request, cutting cost and latency. The key choice — the embedding model and the similarity threshold — governs a hidden trade-off between the hit rate and the false-hit rate (a cached answer served for a different request). Using the public Quora Question Pairs set (40,430 pairs), we compare three compact models (all-MiniLM-L6-v2, e5-small-v2, bge-small-en-v1.5) and two lexical baselines (exact match, TF-IDF) along joint axes: ROC/AUC, hit rate at a fixed false-hit budget, latency and throughput on CPU and GPU, and an estimate of monetary savings. Semantic models (AUC 0.86–0.88) substantially outperform lexical ones (TF-IDF 0.73; exact match catches under 0.5 % of semantic duplicates). bge-small gives the best quality but trails MiniLM in CPU speed; at a strict false-hit budget (1 %) only about one sixth of repeats is separated in the pairwise setting. Embedding cost is negligible relative to the savings on LLM calls. We further show that the operational false-hit rate grows with cache size and exceeds the pairwise estimate. The results give an engineer a basis for choosing a model and threshold for a target profile.

Ключевые слова: семантическое кэширование, большие языковые модели, эмбеддинги предложений, косинусное сходство, ROC-анализ, ложные попадания, оптимизация стоимости.

Keywords: semantic caching, large language models, sentence embeddings, cosine similarity, ROC analysis, false hits, cost optimization.

Введение

Обращения к большим языковым моделям (LLM) через облачные API стали типовым компонентом веб-систем, но каждый вызов стоит денег и добавляет сетевую задержку. На практике, часть пользовательских запросов повторяется по смыслу, отличаясь лишь формулировкой. Семантический кэш перехватывает такой повтор: он кодирует запрос в вектор, ищет ближайший среди ранее обслуженных и, если сходство превышает порог, возвращает готовый ответ, не обращаясь к модели [1]. Такой подход снижает и стоимость, и задержку, и уже реализован в открытых системах (GPTCache [1], MeanCache [6]) и как одна из стратегий сокращения расходов на LLM [3].

Семантический кэш держится на двух устоявшихся компонентах. Первый компонент, векторные представления предложений (от Sentence-BERT [10] до компактных дистиллированных и контрастно обученных моделей MiniLM [13], E5 [12], BGE [14]), кодирует смысл короткого текста для сравнения косинусом. Второй, приближённый поиск ближайшего соседа (FAISS [7], HNSW [8]), обеспечивает масштабирование. Эти компоненты мы не изобретаем, а применяем и сравниваем.

Однако у кэша есть скрытая цена ошибки. При слишком низком пороге сходства кэш вернёт ответ на другой запрос; такое ложное попадание тихо подменяет смысл. Существующие системы [1; 6] фиксируют эмбеддинг-модель де-факто, а универсальные бенчмарки эмбеддеров (MTEB [9], STS [2]) дают агрегатные метрики качества (в том числе на классификации пар), но не измеряют операционный профиль кэша, где на одной оси сходятся доля попаданий при фиксированном бюджете ложных попаданий, выбор порога по ROC и стоимость (latency на CPU/GPU и деньги). Недавние работы затрагивают эту задачу — дообучение доменных эмбеддеров и ансамблирование моделей для кэша [4; 5], — но систематического воспроизводимого сравнения открытых эмбеддеров под семантический кэш, по этим совместным осям, в рецензируемой литературе, насколько нам известно, пока недостаёт; вендорные сравнения встречаются, но не рецензируются и не публикуют воспроизводимую методику.

Мотивация этой работы — практическая. Работая над production-конвейером, который генерирует веб-контент из внешних источников с помощью облачных LLM, мы наблюдали, что одна задача порождает порядка восьмидесяти обращений к модели, и заметная их часть оказывалась почти одинаковой по смыслу, по сути перефразировками. При этом провайдерский prompt-кэш на такой неравномерной нагрузке фактически не срабатывал, а дедупликация выполнялась лишь по точному совпадению ключа, а не по смыслу. Так и возник вопрос, насколько семантическое сопоставление запросов способно перехватить эти повторы и какой ценой.

Цель работы — восполнить этот пробел. Для этого на публичном наборе Quora Question Pairs (QQP) [11] мы сравниваем три компактные эмбеддинг-модели и два лексических базиса, решая задачи: (1) оценить качество разделения дубликатов и не-дубликатов (ROC/AUC); (2) измерить долю попаданий при фиксированном бюджете ложных попаданий; (3) измерить латентность и пропускную способность кодирования на CPU и GPU; (4) измерить латентность поиска в зависимости от размера кэша; (5) оценить денежную экономию. Вклад работы — воспроизводимая методика и количественные ориентиры для выбора модели и порога под целевой профиль нагрузки.

Материалы и методы

Постановка. Пара текстов из QQP имеет метку «дубликат» (вопросы эквивалентны по смыслу) или «не дубликат». Мы трактуем метку как эталон работы кэша: пришёл запрос q_2, в кэше лежит q_1; кэш кодирует оба, считает косинусное сходство s и при $s \geq \theta$ возвращает ответ, связанный с q_1 (попадание). Тогда при пороге θ:

истинное попадание (true hit): $s \geq \theta$ и пара — дубликат (кэш корректно переиспользовал ответ);

ложное попадание (false hit): $s \geq \theta$ и пара — не дубликат (кэш вернул ответ на другой вопрос);

промах (miss): $s < \theta$ (запрос уходит в LLM).

Доля попаданий (hit rate) равна доле верно перехваченных дубликатов (TPR), доля ложных попаданий (false-hit rate) — доле ошибочно перехваченных не-дубликатов (FPR). Изменение θ задаёт ROC-кривую; площадь под ней (AUC) характеризует разделяющую способность модели независимо от выбора порога.

Данные. Набор Quora Question Pairs (в составе GLUE [11]); использован размеченный раздел validation (40 430 пар), из них 14 885 (36,8 %) помечены как дубликаты. Метки взяты как есть, без дополнительной очистки.

Сравниваемые модели. Три компактные эмбеддинг-модели с открытыми весами, все с размерностью вектора 384: all-MiniLM-L6-v2 [13], e5-small-v2 [12] (с рекомендованным префиксом «query:»), bge-small-en-v1.5 [14]. Два лексических базиса: точное совпадение нормализованных строк (нижний регистр, удаление пунктуации) и косинус по TF-IDF-векторам, обученным на всех вопросах набора.

Метрики и стенд. Качество — AUC и рабочие точки (порог и hit rate при бюджете ложных попаданий 1 %, 5 %, 10 %). Латентность кодирования — медиана по трём повторам после прогрева, отдельно на CPU и GPU, на подвыборке 3 000 запросов; отсюда пропускная способность (запросов/с). Абсолютные тайминги сняты на одном десктопе и зависят от фоновой нагрузки, поэтому читаются как индикативные; устойчив здесь относительный порядок, а не точное значение. Латентность поиска — медианное время brute-force поиска ближайшего среди N ранее сохранённых 384-мерных векторов (эмбеддинги bge) при росте N, по 1 000 запросов с прогревом. Модель e5 по рекомендации авторов кодирует оба текста пары с префиксом «query:» (симметричный режим). **Операционная проверка** (модель bge): порог θ калибруется на отдельном calibration-сплите (половина пар) под pairwise-FPR, после чего на независимом test-сплите строится store из N записей и при поиске ближайшего соседа по всему store при росте N измеряются recall и операционный false-hit. Экономия — модель на основе доли повторяемых запросов и recall кэша (см. Результаты). Стенд: Python 3.10, PyTorch 2.11 (CUDA 12.8), sentence-transformers 3.3; CPU и GPU NVIDIA GeForce RTX 5070 Ti; фиксированный seed = 42; код и конфигурации приложены для воспроизводимости. Оценки AUC получены на одном фиксированном разбиении (validation), без доверительных интервалов.

Результаты и обсуждение

Качество разделения

Таблица 1.

Разделяющая способность моделей на QQP (validation)

Модель	Размерность	AUC
bge-small-en-v1.5	384	0,876
all-MiniLM-L6-v2	384	0,871
e5-small-v2	384	0,864
TF-IDF (базис)	—	0,733
точное совпадение (базис)	—	ловит 67 из 14 885 дублей

Три семантические модели близки по AUC (0,864–0,876) и заметно превосходят лексику; TF-IDF даёт 0,733, а точное совпадение строк перехватывает лишь 67 из 14 885 смысловых дубликатов (менее 0,5 %). Смысловые повторы почти никогда не совпадают дословно, поэтому наивный кэш по точному совпадению практически бесполезен, а лексическое сходство (TF-IDF) отстаёт от семантического примерно на 0,14 AUC. На рисунке 1 кривые эмбеддеров лежат заметно выше диагонали и выше TF-IDF, а bge-small стабильно проходит выше остальных.

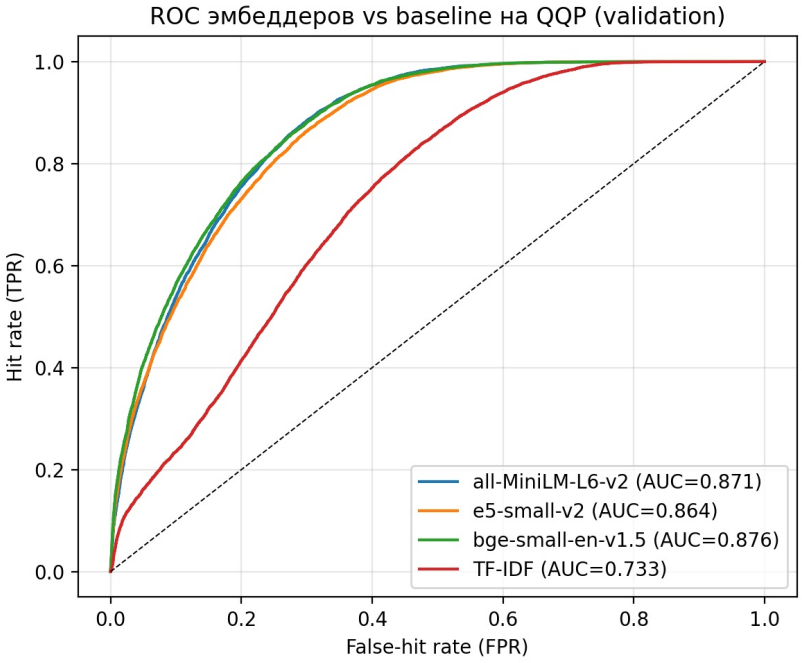

Рисунок 1. ROC-кривые эмбеддеров и базиса TF-IDF на наборе QQP (validation)

Компромисс «попадания против ложных попаданий»

Таблица 2.

Доля попаданий при фиксированном бюджете ложных попаданий

Модель	θ@1 %	hit@1 %	θ@5 %	hit@5 %	θ@10 %	hit@10 %
bge-small-en-v1.5	0,970	0,176	0,930	0,408	0,902	0,564
all-MiniLM-L6-v2	0,965	0,137	0,915	0,359	0,871	0,539
e5-small-v2	0,987	0,153	0,971	0,367	0,959	0,522
TF-IDF	0,943	0,067	0,860	0,162	0,803	0,236

Величины в таблице 2 попарные, то есть false-hit измерен на одной паре. Порог здесь задаёт рабочую точку осознанно, а не «ставится повыше». При попарном бюджете ложных попаданий 1 % даже лучшая модель (bge) отделяет лишь ≈18 % повторов, при 5 % — ≈41 %, и только при 10 % — свыше половины. Безопасность кэша (мало чужих ответов) и его полезность прямо противоречат друг другу, поэтому рабочую точку выбирают под задачу. Модель bge доминирует на всех бюджетах и по качеству предпочтительна. Абсолютные пороги моделей несопоставимы (у e5 рабочие пороги ≈0,96–0,99, у MiniLM ≈0,87–0,97), поэтому порог калибруют под конкретную модель, а не переносят между ними.

Операционная проверка: false-hit растёт с размером кэша

Попарные величины таблицы 2 недооценивают риск реального кэша. В нём запрос сравнивается со всем множеством из N записей, и ложное попадание наступает, если порог превысит хотя бы один не-дубликат (максимум сходства по N), поэтому операционный false-hit при том же пороге обязан расти с размером кэша. Мы проверяем это напрямую на модели bge. Порог θ калиброван на отдельном calibration-сплите под pairwise-FPR 5 % (θ = 0,929), после чего на независимом test-сплите измерены recall и операционный false-hit при поиске ближайшего по store растущего размера.

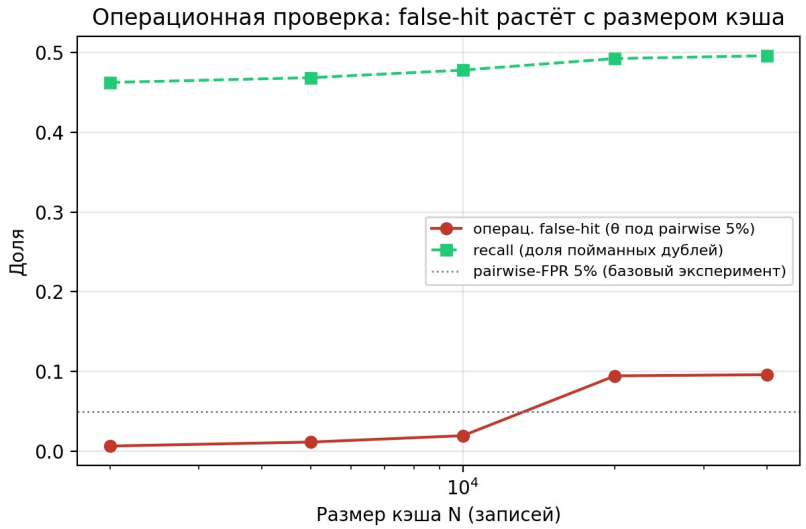

Рисунок 2. Операционный false-hit и recall в зависимости от размера кэша (порог калиброван под pairwise-FPR 5 %)

Механизм подтверждается (рисунок 2). При пороге, дающем 5 % на паре, операционный false-hit растёт с 0,7 % при 2 000 записей до 9,6 % при 40 000 — почти вдвое выше номинального бюджета, тогда как recall держится на уровне 0,46–0,50 (он даже выше попарного, потому что запрос может совпасть с любой записью store, а не только со своей парой). На практике это значит, что попарные пороги таблицы 2 задают лишь нижнюю оценку операционного риска; для большого кэша порог приходится ужесточать под целевой операционный false-hit, а калибровать под реальный размер store.

Стоимость: латентность и пропускная способность

Таблица 3.

Латентность кодирования и пропускная способность (подвыборка 3 000)

Модель	CPU, мс/запрос	CPU, запр./с	GPU, мс/запрос	GPU, запр./с
all-MiniLM-L6-v2	0,74	1343	0,11	9363
bge-small-en-v1.5	1,35	744	0,14	7380
e5-small-v2	1,45	692	0,14	6990

Качество даётся не бесплатно. bge-small на CPU почти вдвое медленнее MiniLM (1,35 против 0,74 мс), а MiniLM при небольшой потере AUC (0,871 против 0,876) даёт наибольшую пропускную способность на CPU. Перенос на GPU (RTX 5070 Ti) ускоряет кодирование примерно в 7–10 раз (MiniLM 0,74 → 0,11 мс, свыше 9 000 запросов/с) и практически снимает стоимость эмбеддинга как узкое место при высокой нагрузке. Значит, выбор модели зависит от бюджета железа. На CPU-хостинге MiniLM даёт лучший компромисс «качество/скорость», а на GPU разрыв между моделями по скорости сжимается, и можно брать более качественную bge.

Латентность поиска и масштаб кэша

Стоимость перебора зависит от размера кэша. На малых кэшах она определяется постоянными накладными расходами (≈7–12 мкс до 1 000 записей), затем растёт с числом записей до ≈400 мкс при 40 000 (рисунок 3). Рост не строго пропорционален N из-за эффектов BLAS и иерархии памяти, но при больших кэшах перебор становится сопоставим со стоимостью самого кодирования и мотивирует переход к приближённому индексу (HNSW [8], FAISS [7]). Для кэшей до нескольких тысяч записей перебор приемлем.

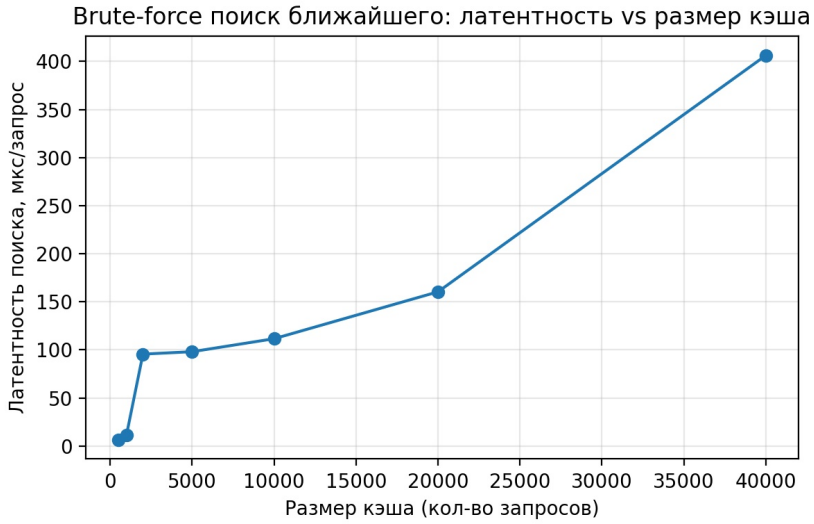

Рисунок 3. Латентность brute-force поиска ближайшего в зависимости от размера кэша

Экономия

Экономию оценим параметрической моделью. Пусть доля повторяемых по смыслу запросов равна r, а recall кэша — доля перехваченных повторов при выбранном бюджете ложных попаданий. Тогда на объём запросов N кэш перехватывает N·r·recall вызовов к LLM — это измеряемая величина, не зависящая от тарифа модели. Значение r зависит от приложения; для ориентира возьмём r ≈ 31 %, сообщённое для реального трафика в MeanCache [6], и покажем диапазон 20–50 %. Модель bge при бюджете ложных попаданий 5 % (recall 0,408) даёт:

Таблица 4.

Доля запросов, обслуженных из кэша (модель bge, recall 0,408), на 1 млн запросов

Повторяемость r	Перехвачено вызовов	Доля всех запросов
20 %	81 540	8,2 %
31 % (MeanCache [6])	126 387	12,6 %
50 %	203 850	20,4 %

В деньгах экономия зависит от тарифа. При публичных ценах API (OpenAI и Anthropic, август 2026 г.) и запросе ≈1000 входных / 500 выходных токенов перехват 126 387 вызовов на 1 млн запросов (r = 31 %) экономит ориентировочно ≈57 $ на GPT-4o-mini, ≈442 $ на Claude Haiku 4.5, ≈948 $ на GPT-4o и ≈1327 $ на Claude Sonnet 4.6; экономия линейна по цене запроса и объёму.

Стоимость самого кодирования пренебрежимо мала относительно вызова LLM (эмбеддинг на CPU — доли миллисекунды локально, вызов LLM — центы плюс сетевой круг), поэтому чистая экономия практически совпадает с валовой. Кэш «почти бесплатен» относительно того, что экономит; ограничителем выступает recall при допустимом уровне ложных попаданий, а не стоимость эмбеддинга.

Заключение

Мы сравнили компактные эмбеддинг-модели для семантического кэширования LLM-запросов по совместному профилю качества, стоимости и экономии на публичном наборе QQP. Семантические модели заметно превосходят лексические базисы, а среди них bge-small даёт лучшее качество на всех бюджетах ложных попаданий, уступая MiniLM в скорости на CPU. Главный практический вывод касается порога сходства. Это осознанный выбор рабочей точки между безопасностью и полезностью кэша (при 1 % ложных попаданий перехватывается лишь ≈1/6 повторов), и калибровать его нужно под конкретную модель. Поскольку стоимость эмбеддинга пренебрежима относительно экономии, основной инженерный рычаг — выбор модели и порога под целевой профиль, а не скорость модели.

Ограничения. QQP — контролируемый прокси; это пары вопросов-парафразов, а не реальный трафик запросов к LLM, поэтому абсолютные пороги и доли повторов в продакшене будут иными; методика (оси и способ выбора рабочей точки) переносится, конкретные числа требуют калибровки на своих данных. Операционный false-hit измерен в предположении, что негативные пробы не имеют неразмеченного дубликата в store; неполнота разметки QQP может приводить к завышению его абсолютного значения, но тренд роста с размером кэша от этого не меняется. Дальнейшая работа — проверка на втором домене данных и сравнение brute-force поиска с приближённым индексом (HNSW/FAISS) по латентности и точности.

Список литературы

1. Bang F. GPTCache: An Open-Source Semantic Cache for LLM Applications Enabling Faster Answers and Cost Savings // Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS). P. 212–218. – Singapore: Association for Computational Linguistics, 2023. – DOI: 10.18653/v1/2023.nlposs-1.24.
2. Cer D., Diab M., Agirre E. [и др.] SemEvalTask 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation // Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval. P. 1–14. – Vancouver: Association for Computational Linguistics, 2017. – DOI: 10.18653/v1/S17-2001.
3. Chen L., Zaharia M., Zou J. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance [Электронный ресурс] // arXiv preprint. – 2023. – DOI: 10.48550/arXiv.2305.05176.
4. Ghaffari S., Bahranifard Z., Akbari M. An Ensemble Embedding Approach for Improving Semantic Caching Performance in LLM-based Systems [Электронный ресурс] // arXiv preprint. – 2025. – DOI: 10.48550/arXiv.2507.07061.
5. Gill W., Cechmanek J., Hutcherson T. [и др.] Advancing Semantic Caching for LLMs with Domain-Specific Embeddings and Synthetic Data [Электронный ресурс] // arXiv preprint. – 2025. – DOI: 10.48550/arXiv.2504.02268.
6. Gill W., Elidrisi M., Kalapatapu P. [и др.] MeanCache: User-Centric Semantic Caching for LLM Web Services [Электронный ресурс] // arXiv preprint. – 2024. – DOI: 10.48550/arXiv. 2403.02694.
7. Johnson J., Douze M., Jégou H. Billion-Scale Similarity Search with GPUs // IEEE Transactions on Big Data. – 2019. – T. 7. – DOI: 10.1109/TBDATA.2019.2921572.
8. Malkov Yu. A., Yashunin D.A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2020. – T. 42. – DOI: 10.1109/TPAMI.2018.2889473.
9. Muennighoff N., Tazi N., Magne L., Reimers N. MTEB: Massive Text Embedding Benchmark // Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL). P. – Dubrovnik: Association for Computational Linguistics, 2037. – DOI: 10.18653/v1/2023.eacl-main.148.
10. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the. P. 3982–3992. – Hong Kong: Association for Computational Linguistics, 2019. – DOI: 10.18653/v1/D19-1410.
11. Wang A., Singh A., Michael J. [и др.] GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding // Proceedings of the. P. 353–355. – Brussels: Association for Computational Linguistics, 2018. – DOI: 10.18653/v1/W18-5446.
12. Wang L., Yang N., Huang X. [и др.] Text Embeddings by Weakly-Supervised Contrastive Pre-training [Электронный ресурс] // arXiv preprint. – 2022. – DOI: 10.48550/arXiv.2212.03533.

13. Wang W., Wei F., Dong L. [и др.] MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers // Advances in Neural Information Processing Systems (NeurIPS). – 2020. – Т. 33.
14. Xiao S., Liu Z., Zhang P., Muennighoff N. C-Pack: Packed Resources For General Chinese Embeddings [Электронный ресурс] // arXiv preprint. – 2023. – DOI: 10.48550/arXiv. 2309.07597.

Статья поступила в редакцию: 14.09.2026
Статья принята к публикации: 20.09.2026
Статья опубликована: 28.09.2026

УДК 004.8; 519.816

DOI: 10.32743/UniTech.2026.150.9.23429

Модульный калибровочный ИИ-рецензент для ранжирования научных статей (Flash-статья)

Кравцов Геннадий Григорьевич
директор
НИЦ Прикладная статистика
Российская Федерация, г. Рязань
E-mail: 62abc@mail.ru

A Modular Calibrating AI Reviewer for Ranking Scientific Papers (Flash Paper)

Kravtsov Gennady Grigorievich
Director of the Research Center Applied Statistics
Russian Federation, Ryazan

Аннотация

Представлена модульная архитектура калибровочного ИИ-рецензента, состоящая из шести независимых модулей, последовательно реализующих определение режима рецензирования (STEM/Humanities), оценку научной новизны, методологическую строгость, практическую ценность, качество визуализации и воспроизводимость. ИИ-рецензент предлагается как инструмент первичной экспресс-фильтрации рукописей, поступающих в редакцию. Валидация архитектуры (DeepSeek-V3, 10 прогонов на 7 статьях) с урезанным средним подтвердила снижение вариативности LLM: коэффициент вариации (CV) сильных работ составил 2–3 %. В ходе инвертированного теста Тьюринга ИИ распознал сгенерированный абсурдный текст, присвоив ему минимальный балл (21.62 из 100) при диагностически высоком CV (28.2 %). Дополнительно в рамках работы апробирован формат Flash-статьи: данная статья-якорь обеспечивает видимость в каталогах и цитируемость, полный текст статьи с дополнительными материалами (логи верификации, код) размещён в репозиториях (SSRN, Zenodo с DOI) и на GitHub. Формат сокращает цикл публикации с 6–12 месяцев до 2–4 недель и предназначен для оперативного представления результатов в журналах с ускоренным рецензированием и печатью, что особенно востребовано в быстроразвивающихся научных областях и прикладных исследованиях.

Библиографическое описание: Кравцов Г.Г. Модульный калибровочный ИИ-рецензент для ранжирования научных статей (Flash-статья) // Universum: технические науки : электрон. научн. журн. 2026. 9(150). URL: https://7universum.com/ru/tech/archive/item/23429

Abstract

A modular architecture of a calibrating AI reviewer is presented, consisting of six independent modules that sequentially implement the determination of the review mode (STEM/Humanities), assessment of scientific novelty, methodological rigor, practical value, visualization quality, and reproducibility. The AI reviewer is proposed as a tool for primary express filtering of manuscripts submitted to an editorial office. Validation of the architecture (DeepSeek-V3, 10 runs on 7 papers) with a trimmed mean confirmed a reduction in LLM variability: the coefficient of variation (CV) for strong papers was 2–3 %. In an inverted Turing test, the AI recognized the generated absurd text, assigning it a minimum score (21.62 out of 100) with a diagnostically high CV (28.2 %). Additionally, the Flash Paper format was tested within the framework of this work: the anchor paper ensures visibility in catalogs and citability, while the full text of the article with supplementary materials (verification logs, code) is deposited in repositories (SSRN, Zenodo with DOI) and on GitHub. The format reduces the publication cycle from 6–12 months to 2–4 weeks and is intended for the prompt presentation of results in journals with accelerated review and printing, which is especially in demand in fast-moving scientific fields and applied research.

Ключевые слова: калибровочный ИИ-рецензент, модульная архитектура, Flash-статья, оперативная публикация, тест Тьюринга.

Keywords: calibrating AI reviewer, modular architecture, Flash Paper, rapid publication, Turing test.

Введение

Актуальность. Экспоненциальный рост научных публикаций [9], и перегрузка экспертов делают необходимыми инструменты автоматизированной предварительной оценки рукописей. Существующие ИИ-рецензенты страдают контекстной зависимостью [3, 4], стохастической нестабильностью [1], они оценивают работу в рамках имеющейся информации [11], тогда как новаторские идеи могут выходить за эти пределы.

Проблема усугубляется издательским лагом: традиционный цикл рецензирования занимает 6–12 месяцев, за которые исследование теряет актуальность. По этой причине в настоящей работе впервые использован формат, названный Flash Paper: препринт представляет полные материалы [7], Github (https://github.com/Famimot/calibrating-ai-reviewer-modular) обеспечивает воспроизводимость результатов, а данная якорная статья с коротким сроком публикации обеспечивает лаконичное и своевременное представление информации.

Простая генерация оценок работы ИИ-рецензента без раскрытия исходных материалов не имеет доказательной силы, а одиночный прогон LLM некорректен в силу стохастической природы модели — необходимы многократные измерения с фильтрацией выбросов.

В прошлой работе [6] предложен монолитный ИИ-рецензент на базе конкурентного ранжирования в замкнутом пакете. Цель работы — спроектировать и валидировать архитектуру из 6 специализированных модулей, отвечающих за отдельный критерий статьи [9, 7]. Валидация проведена на контрастном массиве из 7 статей: 6 реальных публикаций автора с открытым DOI (исключает конфликт интересов) и 1 сгенерированной LLM абсурдной работы (DOI: 10.5281/zenodo.20703374) для инвертированного теста Тьюринга [2].

Это гарантирует полную прозрачность и воспроизводимость, в отличие от исследований на закрытых массивах. Для каждой статьи выполнено 10 прогонов DeepSeek-V3 с расчетом урезанного среднего для фильтрации выбросов.

Бонусный эксперимент проверил способность ИИ распознавать семантический абсурд в формально корректном тексте.

Методология

Принцип калибровки. Рецензент оценивает не отдельную статью по абсолютной шкале, а весь массив одновременно: статьи выступают шкалой друг для друга. Сильная работа получает высокий балл потому, что превосходит остальные, слабая — потому что уступает. Это снимает проблему отсутствия у LLM абсолютной шкалы. Урезанное среднее по 10 прогонам фиксирует устойчивую калибровку, отбрасывая стохастические отклонения.

Архитектура. Выделены фундаментальные сущности оценки научной рукописи, определены их числовые шкалы [5, 8]. Архитектура включает шесть модулей (табл. 1), выполняющих определённую функцию. Это обеспечивает поэтапный контроль корректности, локализацию ошибок и возможность модификации отдельных модулей без переписывания всей системы. Полное описание логики, чек-листов и промптов — в открытом репозитории [7].

Таблица 1.

Модульная структура ИИ-рецензента

Модуль	Название	Функция	Баллы
1	Режим-детектор	Определение STEM/Humanities	0
2	Новизна-анализатор	Оценка новизны, происхождения метода	30
3	Методология-анализатор	Оценка методологической строгости	25
4	Ценность-анализатор	Оценка практической ценности	20
5	Визуализация, воспроизводимость	Оценка визуализации и воспроизводимости	15 + 10
6	Агрегатор-категоризатор	Категоризация, итоговая таблица	100

Процедура рецензирования. Для каждой из R = 10 итераций в «чистом» диалоге с LLM (DeepSeek-V3) последовательно выполняются модули 1–6. Каждый последующий модуль использует результаты предыдущих. Фиксируется итоговая таблица с баллами (0–100) для всех статей. После завершения 10 прогонов исключаются два минимальных и два максимальных значения для каждой статьи, вычисляется урезанное среднее [8]:

$$\overline{S}_i = \frac{1}{R-4} \sum_{r \in R_{trim}} S_i^{(r)}$$

R_{trim} — множество прогонов, исключающих два минимальных и два максимальных значения $S_i^{(r)}$.

R_{trim} — множество прогонов, исключающих два минимальных и два максимальных значения .

Финальный балл рассчитывается как сумма оценок по пяти критериям.

Верификация. Надёжность валидации обеспечивается не количеством статей, а кратным повторением прогонов (R = 10) с фильтрацией выбросов.

Результаты

Ранжирование. После фильтрации получены итоговые баллы для каждой статьи (табл. 2).

Таблица 2.

Результаты 10 прогонов (отфильтрованные данные)

Статья	Новизна (30)	Методология (25)	Практ. Ценность (20)	Визуализация (15)	Воспроизв. (10)	Итог (100)
2	27.83	24.17	17.33	13.67	8.67	91.67
7	27.33	24.17	15.17	12.67	9.5	88.83
5	28.17	24	13.5	12.33	8.83	86.83
3	25	23.83	11.5	12.5	5.83	78.67
1	25	24	11	12.67	5.67	78.33
4	17.5	16.67	6	10.5	5.5	56.17
6	3.8	5.67	2.4	7.5	2.25	21.62

ИИ-рецензент продемонстрировал устойчивую дифференциацию статей. Разрыв между лучшей работой (91.67) и сгенерированной статьёй (21.62) составил 70 баллов. Распределение итоговых баллов до и после фильтрации представлено на рис. 1 (полные данные всех визуализаций — в препринте [7]).

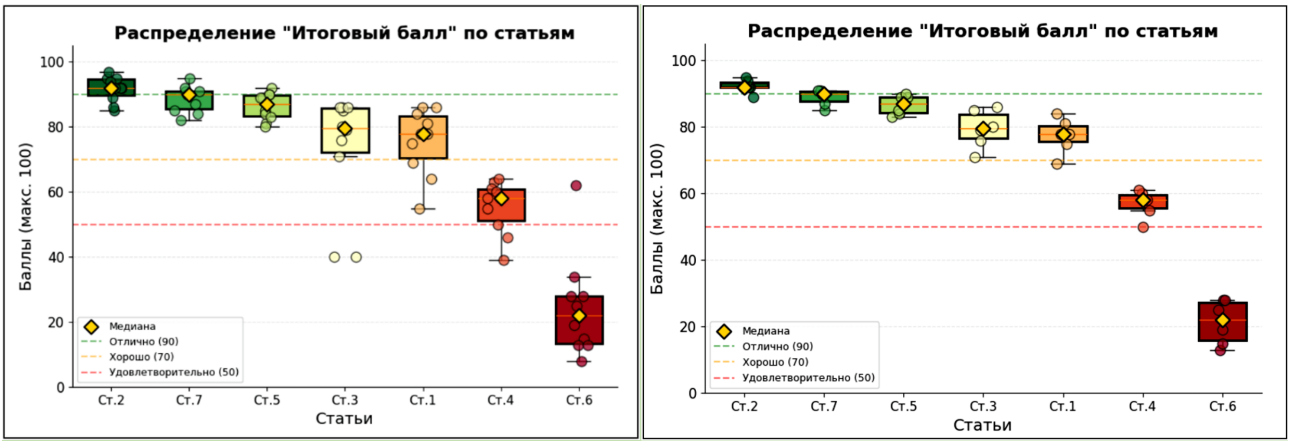

Рисунок 1. Распределение итогового балла: сырые данные (слева) и после фильтрации (справа)

Коэффициент вариации (CV) для сильных статей составляет 2–3 %, для статей средней группы — 6–6.5 %. Для сгенерированной статьи CV = 28.2 % (табл. 3), что более чем в 10 раз превышает CV сильных статей.

Таблица 3.

Вариативность результатов

Paper	Mean	Median	Min	Max	Range	Std	CV (%)
2	92.33	92	89	95	6	1.89	2.04
7	89	90	85	91	6	2.24	2.51
5	86.67	87	83	90	7	2.75	3.17
3	79.5	79.5	71	86	15	5.12	6.44
1	77.5	78	69	84	15	4.72	6.09
4	57	58	50	61	11	3.65	6.41
6	21.33	22	13	28	15	6.02	28.21

Воспроизводимость. Все материалы размещены в открытом доступе: дополнительные материалы (скрипт, верификация) [7],

модульные промпты (рус./англ.),— https://github.com/Famimot/calibrating-ai-reviewer-modular,

текст сгенерированной ИИ статьи — https://doi.org/10.5281/zenodo.20703374,

определение Flash-статьи — https://doi.org/10.5281/zenodo.22727742.

Заключение

Разработанная модульная архитектура подтверждает свою эффективность: формализация критериев в виде модульного конвейера обеспечивает поэтапную локализацию ошибок и возможность модификации отдельных модулей.

Ключевые выводы

1. Калибровочный режим. Оценка отдельной рукописи нестабильна из-за стохастической природы LLM. Предпочтителен режим сравнения, где статьи сами выступают калибровочным материалом.

2. Многократность прогонов. Необходимо выполнять несколько независимых прогонов с фильтрацией экстремумов.

3. Дискриминативная способность. Инструмент устойчиво дифференцирует статьи.

4. Практическая ценность. Инструмент предназначен для первичной фильтрации рукописей и снижения нагрузки на экспертов [3], ранжирования журналов [10], а также самооценки.

Представленный подход демонстрирует полную открытость, воспроизводимость и практическую применяемость в отличие от закрытых систем, не предоставляющих исходных данных для верификации. В то же время использованный формат Flash-статьи позволяет опубликовать работу через несколько недель, а не 6–18 месяцев, без потери актуальности и практической ценности.

Список литературы

1. Bai J. iRULER: Intelligible Rubric-Based User-Defined LLM Evaluation for Revision / J. Bai [et al.] // Proceedings of the. – 2026. – DOI: 10.1145/3772318.3790539.

2. Bhatnagar Y. Breaking the Turing Test: Testing the relevance of the Turing Test against modern LLMs / Y. Bhatnagar // International Journal for Research in Engineering Application & Management. – 2026. – T. 11, № 10.
3. Doskaliuk B. Artificial Intelligence in Peer Review: Enhancing Efficiency While Preserving Integrity / B. Doskaliuk [et al.] // Journal of Korean Medical Science. – 2025. – T. 40, № 7. – DOI: 10.3346/jkms.2025.40.e92.
4. Goyal P. ScholarPeer: A Context-Aware Multi-Agent Framework for Automated Peer Review / P. Goyal [et al.] // arXiv. – 2026. – DOI: 10.48550/arXiv.2601.22638.
5. Institute of Hydrogen Economy. Regulations on the evaluation of scientific works on a 100-point scale based on 10 key criteria [Электронный ресурс]. // International Scientific Journal for Alternative Energy and Ecology (ISJAEE). – 2026. (дата обращения: 13.09.2026).
6. Kravtsov G.G. Two-level open-source architecture for independent AI ranking of scientific journals / G.G. Kravtsov // OSF Preprints. – 2026. – DOI: 10.31235/osf.io/vpbk6_v1.
7. Kravtsov G.G. Калибровочный ИИ-рецензент: модульная архитектура для оценки и ранжирования научных статей / Г.Г. Кравцов. – Research Center Applied Statistics. – 2026. – DOI: 10.5281/zenodo.20732844.
8. Menke J. Establishing institutional scores with the Rigor and Transparency Index: Large-scale analysis of scientific reporting quality / J. Menke [et al.] // Journal of Medical Internet Research. – 2022. – T. 24, № 6. – DOI: 10.2196/37324.
9. Pugliese R. Agentic publications: redesigning scientific publishing in the age of thinking large language models / R. Pugliese [et al.] // Journal of Documentation. – 2026. – T. 82, № 7. – P. 125–149. – DOI: 10.1108/JD-07-2025-0207.
10. Thelwall M. ChatGPT ranking of business and management journals with article quality scores / M. Thelwall // Aslib Journal of Information Management. – 2025. – DOI: 10.1108/AJIM-11-2024-0927.
11. Zhou Y. Can AI Evaluate Literature at the Level of Experts? A Comparative Analysis of ChatGPT-4o, DeepSeek R1, and Human Ratings Based on Entropy / Y. Zhou, H. Hu // Frontiers in Research Metrics and Analytics. – 2025. – T. 10. – Art. 1684137. – DOI: 10.3389/frma.2025.1684137.

References

1. Bai J., Cheong W.S., Muller P., Lim B.Y. iRULER: Intelligible Rubric-Based User-Defined LLM Evaluation for Revision // Proceedings of the. – 2026. – DOI: 10.1145/3772318.3790539.
2. Bhatnagar Y. Breaking the Turing Test: Testing the relevance of the Turing Test against modern LLMs // International Journal for Research in Engineering Application & Management. – 2026. – Vol. 11, No. 10.
3. Doskaliuk B., Zimba O., Yessirkepov M. et al. Artificial Intelligence in Peer Review: Enhancing Efficiency While Preserving Integrity // Journal of Korean Medical Science. – 2025. – Vol. 40, No. 7. – DOI: 10.3346/jkms.2025.40.e92.

4. Goyal P., Parmar M., Song Y., Palangi H., Pfister T., Yoon J. ScholarPeer: A Context-Aware Multi-Agent Framework for Automated Peer Review // arXiv. – 2026. – DOI: 10.48550/arXiv.2601.22638.
5. Institute of Hydrogen Economy. Regulations on the evaluation of scientific works on a 100-point scale based on 10 key criteria // International Scientific Journal for Alternative Energy and Ecology (ISJAEE). – 2024. – URL: https://www.isjaee.com/jour/announcement/view/446 (дата обращения: 01.04.2026).
6. Kravtsov G.G. Two-level open-source architecture for independent AI ranking of scientific journals // OSF Preprints. – 2026. – DOI: 10.31235/osf.io/vpbk6_v1.
7. Kravtsov G. Calibrating AI Reviewer: Modular architecture for evaluating and ranking scientific articles. Research Center Applied Statistics. – 2026. – DOI: 10.5281/zenodo.20732844.
8. Menke J., Eckmann P., Ozyurt I.B., Roelandse M., Anderson N., Grethe J., Gamst A., Bandrowski A. Establishing institutional scores with the Rigor and Transparency Index: Large-scale analysis of scientific reporting quality // Journal of Medical Internet Research. – 2022. – Vol. 24, No. 6. – DOI: 10.2196/37324.
9. Pugliese R., Kourousias G., Venier F., Garlatti Costa G. Agentic publications: redesigning scientific publishing in the age of thinking large language models // Journal of Documentation. – 2026. – Vol. 82, No. 7. – DOI: 10.1108/JD-07-2025-0207.
10. Thelwall M. ChatGPT ranking of business and management journals with article quality scores // Aslib Journal of Information Management. – 2025. – DOI: 10.1108/AJIM-11-2024-0927.
11. Zhou Y., Hu H. Can AI Evaluate Literature at the Level of Experts? A Comparative Analysis of ChatGPT-4o, DeepSeek R1, and Human Ratings Based on Entropy // Frontiers in Research Metrics and Analytics. – 2025. – Vol. 10. – Art. 1684137. – DOI: 10.3389/frma.2025.1684137.

Статья поступила в редакцию: 29.08.2026
Статья принята к публикации: 05.09.2026
Статья опубликована: 28.09.2026

УДК 004.89+006.3

DOI: 10.32743/UniTech.2026.150.9.23376

Оптимизация архитектуры интеллектуального ассистента технической документации на базе LLM без использования векторных баз данных

Сперанский Роман Александрович
специалист по информационной безопасности, старший инженер-программист, независимый исследователь, McAfee Corp
Канада, г. Ванкувер
E-mail: r.speranskii@gmail.com

Optimization of technical documentation smart assistant architecture based on LLM without using vector databases

Speranskii Roman
Information Security Specialist, Senior Software Engineer, Independent Researcher, McAfee Corp,
Canada, Vancouver

Аннотация

Цель исследования – обоснование и производственная апробация экономически эффективной архитектуры диалогового ассистента по продуктовой технической документации малого объёма (порядка 5 000 – 20 000 токенов) без применения Retrieval-Augmented Generation (RAG) и векторных баз данных. Методология: спроектирована и внедрена в производственную среду архитектура прямой инжекции статического контекста (Direct Context Injection) с кэшированием префикса промпта (Prompt Caching), ролевым разграничением доступа на уровне промпта и изоляцией модели от внешних инструментов; выполнено аналитическое сравнение с классическим RAG-конвейером по составу инфраструктуры, этапам обработки запроса и режимам отказов; построена расчётная модель стоимости эксплуатации на открытых тарифах LLM-провайдера. Результаты: паттерн исключает конвейер векторизации и векторное хранилище, устраняет по построению класс ошибок неполного извлечения контекста (retrieval miss), обеспечивает контролируемое обновление базы знаний по схеме human-in-the-loop; расчётная стоимость запроса при контексте 15 000 токенов и прогретом кэше – около 0,003 долл. Вывод и практическая значимость: для справочных корпусов малого объёма RAG избыточен; паттерн, критерий

Библиографическое описание: Сперанский Р.А. Оптимизация архитектуры интеллектуального ассистента технической документации на базе LLM без использования векторных баз данных // Universum: технические науки : электрон. научн. журн. 2026. 9(150). URL: https://7universum.com/ru/tech/archive/item/23376

выбора архитектуры и модель стоимости применимы разработчиками B2B- и B2C-сервисов для бюджетного внедрения ИИ-ассистентов. Экспериментальное сравнение качества подходов на едином тестовом наборе – направление дальнейшей работы.

Abstract

Purpose of the study: justification and production validation of a cost-effective architecture for a conversational assistant over small-volume product technical documentation (about 5,000-20,000 tokens) without Retrieval-Augmented Generation (RAG) and vector databases. Methodology: a Direct Context Injection architecture with prompt prefix caching, prompt-level role-based access control, and tool-less model isolation was designed and deployed to production; an analytical comparison with the classical RAG pipeline was performed in terms of infrastructure, request processing stages, and failure modes; a verifiable operational cost model was built on the provider's published tariffs. Results: the pattern eliminates the vectorization pipeline and the vector store, removes the retrieval-miss error class by construction, and provides controlled human-in-the-loop knowledge base updates; the calculated cost of a request with a 15,000-token context and a warm cache is about $0.003. Conclusion and practical significance: for small reference corpora RAG is redundant; the pattern, the corpus-size selection criterion, and the cost model can be adopted by B2B and B2C vendors for rapid, low-cost deployment of AI assistants. Experimental quality comparison on a unified test set is identified as future work.

Ключевые слова: большие языковые модели, RAG, векторные базы данных, кэширование промптов, прямая инжекция контекста, ролевое разграничение доступа, стоимость эксплуатации.

Keywords: large language models, RAG, vector databases, prompt caching, direct context injection, role-based access control, operational cost.

Введение

Интеграция ассистентов на базе больших языковых моделей (LLM) в пользовательские интерфейсы программных продуктов стала стандартом де-факто. Доминирующий паттерн ответов на вопросы по закрытым данным - Retrieval-Augmented Generation (RAG) [5; 8]: корпус фрагментируется (chunking), фрагменты векторизуются и сохраняются в векторной базе данных, при запросе выполняется семантический поиск k релевантных фрагментов, из которых собирается контекст генерации.

Для узкой задачи – ассистента по документации одного продукта малого масштаба – RAG часто избыточен: он требует сопровождения векторного хранилища и конвейера переиндексации и вносит характерный режим отказа – потерю релевантного контекста при извлечении (retrieval miss). Рост контекстных окон LLM, их способность решать задачи по предоставленному контексту без дообучения [4] и штатные механизмы кэширования префикса промпта [3; 6] актуализировали альтернативу – прямую инжекцию полного корпуса в контекст модели (Direct Context Injection, также Context Stuffing). По данным сравнительных исследований, при достаточном бюджете контекста длинноконтекстная генерация в среднем превосходит RAG по качеству, преимуществом RAG остаётся стоимость обработки больших корпусов [9]; известное ограничение длинных контекстов –

неравномерное использование информации в зависимости от позиции («lost in the middle») [10] – мотивирует ограничить область применения прямой инжекции корпусами малого объёма.

Цель исследования - обоснование и производственная апробация паттерна интеграции LLM-ассистента документации без векторных баз данных, объединяющего прямую инжекцию статического контекста, кэширование префикса промпта, ролевое разграничение доступа на уровне промпта и изоляцию модели от внешних инструментов. Задачи: 1) описать архитектуру и её производственную реализацию; 2) выполнить аналитическое сравнение с классическим RAG-конвейером; 3) построить верифицируемую модель стоимости; 4) определить границы применимости и ограничения выводов.

Отдельные элементы решения - длинноконтекстная генерация, кэширование промптов, ролевая модель доступа, отказ от инструментов - известны. Научно-практический результат работы состоит не в новом алгоритме, а в следующем: (а) систематизация этих механизмов в воспроизводимый паттерн для класса задач «ассистент по документации малого объёма»; (б) явный критерий выбора между RAG и прямой инжекцией по объёму корпуса и профилю нагрузки; (в) верифицируемая модель стоимости эксплуатации на открытых тарифах; (г) схема контролируемого обновления базы знаний по принципу human-in-the-loop.

Материалы и методы

Объект исследования. Объектом исследования является справочная подсистема производственного веб-приложения для управления недвижимостью (далее - Acme Property Management; наименование обезличено). База знаний – единый файл help.md в формате Markdown объёмом порядка 15 000 токенов, что типично для справочных систем узкоспециализированных B2B-продуктов (5 000 – 20 000 токенов). Файл состоит из разделов; каждый описывает одну пользовательскую функцию (бронирование объектов, создание здания, регистрация заявки на обслуживание) и имеет единообразную структуру: заголовок, пошаговая инструкция и ролевая аннотация «Who can: …», задающая на естественном языке перечень ролей (resident, manager, admin), которым доступно действие. При добавлении или изменении функциональности соответствующий раздел редактируется вручную; отдельный конвейер преобразования данных отсутствует – это принципиальное свойство решения.

Метод исследования

Работа выполнена как инженерное исследование в форме производственного кейса в сочетании с аналитическим сравнением архитектур. Решение реализовано на двух стеках (TypeScript, Node.js/Express и Kotlin/Ktor со встроенным HTTP-клиентом JDK) и развёрнуто в производственной среде; генерация выполняется моделью claude-haiku-4-5 через API Anthropic [1]. Сравнение с классическим RAG-конвейером (чанкинг → векторизация → векторная БД → извлечение top-k → сборка промпта) проведено аналитически: по составу инфраструктуры, этапам обработки запроса, режимам отказов и стоимости, с опорой на опубликованные сравнительные исследования [9; 10]. Собственное экспериментальное измерение точности и частоты галлюцинаций двух подходов на едином тестовом наборе не

проводилось и определено как направление дальнейшей работы; количественные утверждения о качестве ограничены результатами цитируемых работ, а собственные количественные результаты статьи относятся к модели стоимости.

Обеспечение изоляции

Изоляция модели от внешних источников реализована на уровне протокола взаимодействия с LLM API: в запросах не передаётся массив tools (function calling), поэтому у модели нет канала для обращения к внешним ресурсам или исполнения кода – её возможности ограничены генерацией текста по переданному контексту. Это программная, а не аппаратная мера: исходящие сетевые вызовы выполняет только серверный компонент к единственной конечной точке LLM API, а ограничение исходящего трафика (egress-политики) остаётся задачей инфраструктурного контура и в настоящей работе не рассматривается. Ассистент вынесен за пределы доверенной зоны: он не выполняет действий в системе, а только формирует справочные ответы; фактическое разграничение прав обеспечивает основной API приложения. Это согласуется с рекомендациями по снижению рисков LLM-приложений [11], в том числе для косвенных инъекций промптов [7]: даже при успешной инъекции через текст вопроса модель не располагает ни инструментами, ни полномочиями в системе.

Результаты и обсуждение

Предлагаемая архитектура

Архитектура сведена к трём логическим узлам (рис. 1): база знаний help.md; серверный слой, выполняющий аутентификацию (JWT), сборку системного сообщения и проксирование запроса к LLM API (ключ API существует только в конфигурации сервера и недоступен клиенту); клиентский чат-виджет, взаимодействующий исключительно с собственным серверным компонентом.

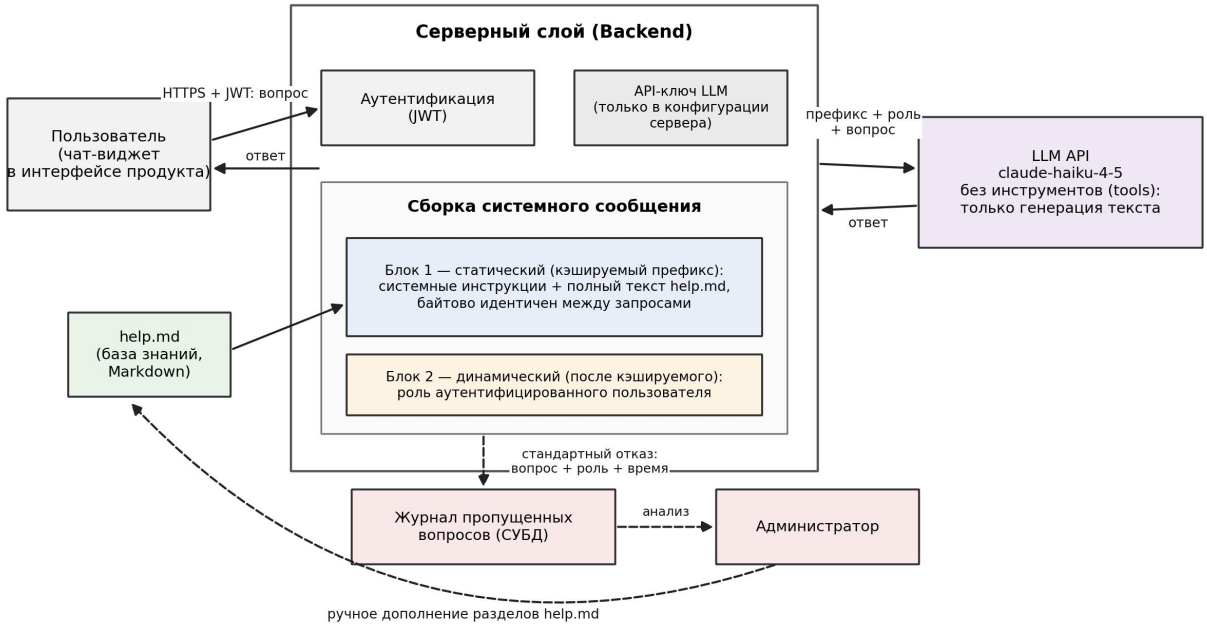

Рисунок 1. Предлагаемая архитектура: формирование системного сообщения, кэшируемый префикс и контур обновления базы знаний

Системное сообщение разделено на две части (табл. 1). Статическая часть - системные инструкции и полный текст документации - помечается атрибутом cache_control: {type: «ephemeral»} и остаётся байтово идентичной между запросами, благодаря чему провайдер кэширует обработанный префикс и повторно использует его в последующих вызовах [3]. Динамическая часть - роль аутентифицированного пользователя - размещается строго после кэшируемого блока и в кэш не попадает. Попадание любых волатильных данных (временных меток, идентификаторов запроса) в статический блок приводит к «тихим» промахам кэша; корректность контролируется по полям usage ответа API (cache_creation_input_tokens, cache_read_input_tokens).

Таблица 1.

Структура блоков системного сообщения при вызове LLM API

№	Назначение блока	Изменяемость	Кэширование
1	Системные инструкции и полный текст документации (<docs>…</docs>)	Статический; байтово идентичен между запросами	cache_control: {type: "ephemeral"} - кэшируемый префикс
2	Роль аутентифицированного пользователя	Изменяется от запроса к запросу; размещается строго после кэшируемого блока	Не кэшируется

Листинг 1 демонстрирует формирование тела запроса на языке Kotlin; переменная systemPrompt содержит инструкции и полный текст help.md, загружаемый однократно при старте сервиса.

```
val payload = AnthropicRequest(
model = «claude-haiku-4-5»,
maxTokens = 1024,
```

```
system = listOf(
SystemBlock(
text = systemPrompt,
cacheControl = CacheControl(type = «ephemeral»)
),
SystemBlock(text = «The person asking has the role: $role.»)
),
messages = listOf(Message(«user», question))
)
```

Ролевое разграничение реализовано так: системные инструкции предписывают модели отвечать только по документации, отказываться от нерелевантных запросов и не описывать шаги, недоступные роли спрашивающего; носителем модели доступа выступают аннотации «Who can:» в самой документации, что исключает дублирование матрицы прав в коде ассистента. Граница применимости: это эргономическая мера (не подсказывать пользователю недоступные ему действия), а не механизм безопасности - фактическую авторизацию операций выполняет основной API продукта.

Сравнение с RAG-конвейером

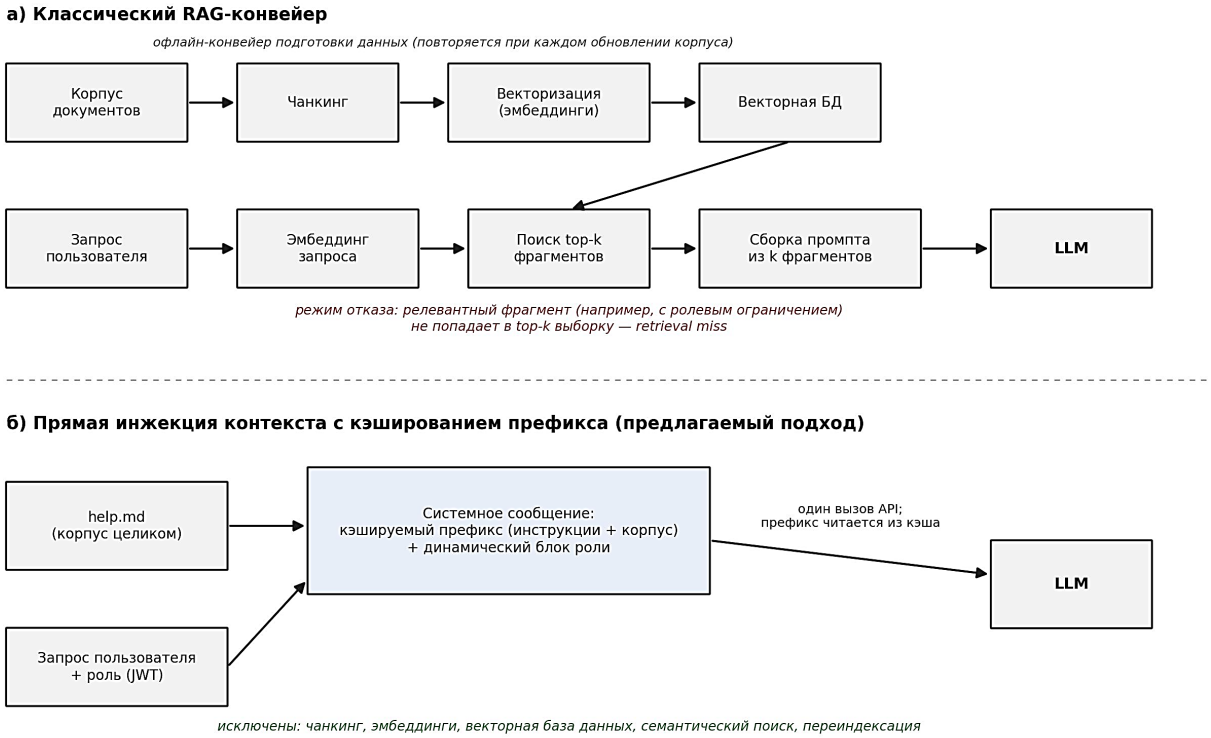

Рисунок 2. Сравнение конвейеров обработки запроса: классический RAG (а) и прямая инжекция контекста с кэшированием префикса (б)

Сравнительная матрица этапов обработки запроса приведена в таблице 2. Прямая инжекция устраняет по построению класс ошибок неполного извлечения (retrieval miss), при котором релевантный фрагмент - например, содержащий ролевое ограничение - не попадает в top-k выборку: модель всегда получает корпус целиком, что согласуется с результатами [9]. Вместе с тем полный контекст сам по себе не гарантирует полного использования

содержащейся в нём информации [10] и не исключает генеративных галлюцинаций; инструктивное ограничение (отвечать только по документации, отказ по умолчанию), отсутствие внешних источников и малый объём корпуса снижают риски, однако количественная оценка частоты галлюцинаций требует отдельного эксперимента (см. подраздел «Ограничения»).

Таблица 2.

Сравнительная матрица обработки запроса

Этап	Классический RAG	Предлагаемая архитектура
Подготовка данных	Чанкинг → векторизация → загрузка в векторную БД; переиндексация при каждом обновлении корпуса	Однократная сборка файла help.md; обновление - правка раздела файла
Клиентский слой	HTTP-запрос с JWT	HTTP-запрос с JWT
Серверный слой	Векторизация запроса → поиск top-k → сборка промпта из фрагментов	Агрегация статического префикса и роли пользователя
Вызов LLM API	Генерация на основе k извлечённых фрагментов	Чтение кэшированного префикса (~15 тыс. токенов) → генерация

Разграничение авторского вклада

Штатные механизмы, используемые без модификации: языковая модель и её API [1], кэширование префикса промпта [3], аутентификация JWT. Авторские проектные решения: 1) структура системного сообщения с разделением на кэшируемый статический и некэшируемый ролевой блоки и дисциплина байтовой стабильности префикса; 2) аннотации «Who can:» внутри Markdown-документации как единственный носитель модели доступа ассистента; 3) tool-less конфигурация как осознанная мера изоляции модели; 4) схема контролируемого обновления базы знаний с журналированием отказов; 5) критерий выбора архитектуры по объёму корпуса. Совокупность этих решений и составляет предлагаемый паттерн; утверждений о создании нового алгоритма семейства RAG или нового механизма кэширования работа не содержит.

Модель стоимости эксплуатации

Расчёт выполнен по опубликованным тарифам claude-haiku-4-5 [2] на дату обращения: базовые входные токены - 1,00 долл. за 1 млн; выходные - 5,00; запись кэша (TTL 5 минут) - 1,25; чтение кэша - 0,10. Сценарий: кэшируемый префикс (инструкции и документация) - 15 000 токенов; средняя длина ответа - 300 токенов; вопрос пользователя мал по сравнению с префиксом и в оценке опущен; кэш прогрет - интервалы между запросами не превышают TTL, продлеваемого при каждом обращении, что типично для частотного пользовательского трафика.

Стоимость одного запроса при прогретом кэше: $C = 15\,000 / 10^6 \times 0{,}10$ долл. $+ 300 / 10^6 \times 5{,}00$ долл. $= 0{,}0015 + 0{,}0015 = 0{,}003$ долл. Первый («холодный») запрос дополнительно оплачивает запись кэша: $15\,000 / 10^6 \times 1{,}25$ долл. $+ 0{,}0015$ долл. $\approx 0{,}020$ долл. Сравнение режимов на горизонте 1 000 запросов приведено в таблице 3; в отличие от предыдущих редакций расчёта, входные токены при кэшировании учитываются в полном объёме - они продолжают передаваться, но тарифицируются по ставке чтения кэша.

Таблица 3.

Расчётная стоимость обработки 1 000 запросов (префикс 15 000 токенов, ответ 300 токенов, тарифы claude-haiku-4-5 [2])

Режим	Входные токены на запрос и тарификация	Вход, долл.	Генерация, долл.	Итого, долл.
Прямая инжекция без кэширования	15 000; базовый тариф, 1,00 долл./1 млн	15,00	1,50	≈ 16,50
Прямая инжекция с кэшированием (предлагаемый)	15 000; чтение кэша, 0,10 долл./1 млн	1,50	1,50	≈ 3,00 *
RAG, top-3 фрагмента (аналитическая оценка)	≈ 2 000; базовый тариф, 1,00 долл./1 млн	2,00	1,50	≈ 3,50 **

* дополнительно ≈ 0,02 долл. - однократная запись кэша при холодном запросе (15 000 × 1,25 долл./1 млн); ** без учёта затрат на векторизацию корпуса и запросов, а также постоянных затрат на инфраструктуру векторной базы данных и сопровождение переиндексации.

Из таблицы 3 следует, что при малых корпусах токенные затраты подходов сопоставимы: преимущество паттерна создаётся не тарифами на генерацию, а устранением инфраструктурных компонентов (векторная БД, сервис эмбеддингов) и этапов сопровождения (переиндексация). Показатель 0,003 долл. не является универсальной константой метода: он линейно зависит от объёма префикса (при 100 000 токенов чтение кэша даёт 0,010 долл., итого ≈ 0,0115 долл. за запрос), от длины ответа и доли попаданий в кэш; при разреженном трафике с интервалами больше TTL стоимость смещается к 0,020 долл. за запрос. Все составляющие расчёта верифицируемы по полям usage ответов API и открытым тарифам [2; 3].

Контролируемое обновление базы знаний

Вместо автономного «самообучения» бота, поведение которого не поддаётся аудиту, реализована пассивная петля обратной связи по принципу human-in-the-loop. Если ответ ассистента - стандартный отказ («могу помочь только по вопросам продукта…»), серверный компонент фиксирует вопрос, роль пользователя и время в реляционной БД; администратор периодически анализирует журнал пропущенных вопросов и вручную дополняет разделы help.md, а обновлённый файл автоматически образует новый кэшируемый префикс при следующем запросе. Каждое изменение проходит через человека и фиксируется в истории файла - обновление базы знаний контролируемо и трассируемо. Детерминированной генерацию ответов это не делает: стохастичность выборки токенов свойственна самой языковой модели; корректной характеристикой схемы является контролируемость обновления знаний, а не детерминированность ответов.

Ограничения

Ограничениями настоящей работы являются: 1) один производственный продукт и один провайдер LLM - выводы о стоимости зависят от тарифной политики провайдера и подлежат пересчёту при её изменении; 2) отсутствие собственного экспериментального сравнения качества RAG и прямой инжекции на едином тестовом наборе - утверждения о качестве

опираются на цитируемые внешние исследования [9; 10]; 3) выводы ограничены фактически исследованным диапазоном объёмов корпуса (порядка 5 000-20 000 токенов). Экстраполяция на корпуса в 50 000-100 000 токенов следует из линейности модели стоимости и результатов [9], однако до экспериментальной проверки должна рассматриваться как гипотеза. Дальнейшая работа: открытый тестовый набор запросов трёх классов (по интерфейсу, нерелевантные, с нарушением ролевой модели), референсный RAG-контур на том же корпусе и измерение точности, частоты галлюцинаций и задержки обоих подходов.

Заключение

Для справочных корпусов малого объёма (порядка 5 000-20 000 токенов) классическая RAG-архитектура избыточна: паттерн прямой инжекции статического контекста с кэшированием префикса воспроизводит целевую функциональность ассистента без конвейера подготовки данных и дополнительной инфраструктуры, устраняет по построению ошибки неполного извлечения контекста и обеспечивает расчётную стоимость запроса около 0,003 долл. в описанных условиях. Ролевое разграничение на уровне промпта и tool-less конфигурация формируют консервативный профиль безопасности, а петля human-in-the-loop - контролируемое обновление базы знаний.

Предложенный паттерн, критерий выбора архитектуры по объёму корпуса и верифицируемая модель стоимости непосредственно применимы разработчиками B2B- и B2C-сервисов для быстрого и бюджетного внедрения ассистентов по документации. Количественное сравнение качества с RAG на едином тестовом наборе и уточнение верхней границы применимости паттерна по объёму корпуса - предмет дальнейших исследований.

Список литературы

1. Anthropic PBC. Introducing Claude Haiku 4.5. – 2025.обращения: 31.08.2026) [Электронный ресурс]. URL: https://www.anthropic.com/news/claude-haiku-4-5(дата. (дата обращения 24.09.2026)
2. Anthropic PBC. Pricing. Claude Developer Documentation [Электронный ресурс]. – 2026. URL: https://docs.anthropic.com/en/docs/about-claude/pricing (дата обращения: 31.08.2026).
3. Anthropic PBC. Prompt caching. Claude Developer Documentation.обращения: 31.08.2026) [Электронный ресурс]. URL: https://docs.anthropic.com/en/docs/build-with-claude/prompt-caching(дата. (дата обращения 24.09.2026)
4. Brown T.B., Mann B., Ryder N. et al. Language models are few-shot learners // Advances in Neural Information Processing Systems. – 1901. – T. 33. – P. 1877–1901.
5. Gao Y., Xiong Y., Gao X. et al. Retrieval-augmented generation for large language models: a survey // arXiv preprint arXiv:2312.10997. – 2023. – DOI: 10.48550/arXiv.2312.10997 [Электронный ресурс]. URL: https://arxiv.org/abs/2312.10997 (дата обращения: 31.08.2026).
6. Gim I., Chen G., Lee S., Sarda N., Khandelwal A., Zhong L. Prompt cache: modular attention reuse for low-latency inference // Proceedings of Machine Learning and Systems. – 2024. – T. 6. – P. 325–338.

7. Greshake K., Abdelnabi S., Mishra S., Endres C., Holz T., Fritz M. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection // Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23). – 2023. – P. 79–90. – DOI: 10.1145/3605764.3623985.
8. Lewis P., Perez E., Piktus A. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks // Advances in Neural Information Processing Systems. – 2020. – T. 33. – P. 9459–9474.
9. Li Z., Li C., Zhang M., Mei Q., Bendersky M. Retrieval augmented generation or long-context LLMs? A comprehensive study and hybrid approach // Proceedings of the. – 2024. – P. 881–893. – DOI: 10.18653/v1/2024.emnlp-industry.66.
10. Liu N.F., Lin K., Hewitt J. et al. Lost in the middle: how language models use long contexts // Transactions of the Association for Computational Linguistics. – 2024. – T. 12. – P. 157–173. – DOI: 10.1162/tacl_a_00638.
11. OWASP Foundation. OWASP Top 10 for Large Language Model Applications: 2025 [Электронный ресурс]. – 2024. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/ (дата обращения: 31.08.2026).

ЭЛЕКТРОНИКА, ПРИБОРОСТРОЕНИЕ И РАДИОТЕХНИКА

Статья поступила в редакцию: 08.08.2026
Статья принята к публикации: 15.08.2026
Статья опубликована: 28.09.2026

УДК 535.61+665.6+621.383

DOI: 10.32743/UniTech.2026.150.9.23363

Математическая модель прохождения инфракрасного луча для нефтепродукции

Алиев Ибратжон Хатамович
директор
Научно-исследовательский институт Физики резонансных ядерных реакций
Республика Узбекистан, г. Маргилан

Эргашев Шохбозжон Умарали угли
соискатель
Ферганский государственный технический университет
Республика Узбекистан, г. Фергана
E-mail: almaxanparpiyeva@gmail.com

Mathematical model of infrared beam propagation for petroleum products

Aliyev Ibratjon Khatamovich
Director
Research Institute of Resonance Nuclear Reaction Physics
Republic of Uzbekistan, Margilan

Ergashev Shokhbozjon Umarali ugli
Independent Researcher
Fergana State Technical University
Republic of Uzbekistan, Fergana

Аннотация

Работа посвящена построению математической модели оптического датчика влажности бензина, предназначенного для контроля обводнённости топлива в баках автомобилей специального назначения (машин скорой медицинской помощи, пожарных автомобилей), эксплуатируемых в условиях жаркого климата. Принцип действия датчика основан на измерении ослабления инфракрасного излучения при прохождении через слой топлива между источником излучения и фоторезистором. В качестве базового физического закона использован закон Бугера–Ламберта–Бера, обоснован выбор данного приближения в

Библиографическое описание: Алиев И.Х., Эргашев Ш.У.У. Математическая модель прохождения инфракрасного луча для нефтепродукции // Universum: технические науки : электрон. научн. журн. 2026. 9(150). URL: https://7universum.com/ru/tech/archive/item/23363

сравнении с более общими моделями переноса излучения и теорией Ми. Получено общее решение дифференциального уравнения ослабления интенсивности, введена линейная калибровочная зависимость коэффициента ослабления от концентрации воды в топливе, на основании модельных калибровочных данных решена система уравнений относительно калибровочных параметров и построена частная форма модели. Приведены результаты численного моделирования зависимости пропускания излучения от концентрации воды и от длины зазора между источником и приёмником в диапазоне, соответствующем реальным геометрическим размерам горловин топливных баков. Результаты моделирования могут быть использованы при разработке бортовой аппаратуры непрерывного контроля степени влажности топлива, эксплуатируемой как в статическом, так и в динамическом режиме работы автомобиля.

Abstract

This study is devoted to the development of a mathematical model for an optical gasoline moisture sensor designed for monitoring fuel water content in the fuel tanks of special-purpose vehicles (such as ambulances and fire engines) operating in hot climate conditions. The operating principle of the sensor is based on measuring the attenuation of infrared radiation as it passes through a fuel layer positioned between an emitter and a photoresistor.

The Beer–Lambert–Beer law serves as the core physical framework, and the selection of this approximation is justified in comparison with more general radiation transport models and Mie scattering theory. The general solution to the differential equation governing intensity attenuation was derived, introducing a linear calibration dependence of the attenuation coefficient on the water concentration in the fuel. Based on model calibration data, the system of equations was solved relative to the calibration parameters, establishing a specific form of the model.

Numerical simulation results demonstrate the dependence of radiation transmittance on water concentration and on the gap distance between the emitter and the receiver within range bounds corresponding to the actual geometric dimensions of fuel tank filler necks. The modeling results can be applied to the development of on-board equipment for continuous monitoring of fuel moisture levels, suitable for operation in both static and dynamic vehicle operating modes.

Ключевые слова: закон Бугера–Ламберта–Бера, инфракрасный датчик, влажность бензина, степень влажности топлива, фоторезистор, математическое моделирование, специальный автотранспорт.

Keywords: Beer–Lambert–Beer law, infrared sensor, gasoline moisture, fuel moisture degree, photoresistor, mathematical modeling, special vehicles.

Введение

Степень влажности жидкого моторного топлива является одним из ключевых параметров, определяющих его эксплуатационное качество и надёжность работы двигателя внутреннего сгорания. В том числе особенно следует подчеркнуть важность данной тематики в отношении специализированной и военной техники. Присутствие в бензине свободной или растворённой воды [1; 9; 15] приводит к нарушению стабильности горения топливно-

воздушной смеси, коррозии элементов топливной системы, а при низких температурах — к образованию ледяных пробок в топливопроводах. Отдельную и практически значимую проблему представляет собой обводнение топлива в баках автомобилей специального назначения — машин скорой медицинской помощи, пожарных автомобилей, военной техники, для которых отказ двигателя при экстренном вызове напрямую связан с риском для жизни. Более того, актуальной проблематикой для описываемых видов транспортных средств аналогично остаётся функционирование в экстремальных полевых условиях, на что стоит уделять особенное внимание.

Появление воды в топливном баке в значительной степени обусловлено явлением так называемого «дыхания» бака: суточные колебания температуры вызывают попеременное расширение и сжатие паровоздушной смеси в над-топливном пространстве, вследствие чего при остывании бака в него засасывается атмосферный воздух с последующей конденсацией содержащейся в нём влаги. В условиях жаркого климата с большими суточными перепадами температур, характерными для Республики Узбекистан, интенсивность подобных термоциклов существенно возрастает, что делает задачу оперативного контроля влажности топлива особенно актуальной для местного специального автотранспорта.

Существующие методы определения содержания воды в топливе — от лабораторного титрования по методу Карла Фишера до импедансной спектроскопии [11] — обеспечивают высокую точность, но требуют либо отбора и транспортировки пробы, либо стационарного лабораторного оборудования, что делает их непригодными для непрерывного бортового контроля [6; 13]. В последние годы активно развиваются портативные и встраиваемые оптические методы контроля качества топлив, основанные на спектроскопии в ближней инфракрасной области, позволяющие проводить измерения непосредственно в месте эксплуатации в реальном времени [6; 14]. Отдельное направление составляют недорогие оптические датчики мутности (турбидиметры и нефелометры [8]) на основе пары «инфракрасный излучатель — фотоприёмник», первоначально разработанные для контроля качества природных и сточных вод, но по своей физической природе применимые и к задаче обнаружения свободной и эмульгированной воды в углеводородных средах [2; 7].

Целью настоящей работы является построение математической модели простейшего оптического датчика влажности бензина, использующего источник инфракрасного излучения и фоторезистор, расположенные по разные стороны от контролируемого объёма топлива. В качестве базового приближения выбран закон Бугера–Ламберта–Бера [3; 4; 10; 12; 16; 17] как наиболее простая и физически обоснованная модель ослабления излучения в среде с малой и умеренной концентрацией растворённой примеси, исходя из всего вышеперечисленного.

Материалы и методы

Общей теоретической основой для описания взаимодействия оптического излучения с веществом служит уравнение переноса излучения, учитывающее одновременно поглощение, испускание и многократное рассеяние излучения по всем направлениям в среде. Его точное решение в общем виде требует численных методов и оправдано для сред с сильным многократным рассеянием — эмульсий высокой концентрации, биологических тканей, мутных природных вод. Частным и значительно более простым случаем данного

уравнения является закон Бугера–Ламберта–Бера, справедливый при следующих условиях: излучение является коллимированным, взаимодействие сводится преимущественно к поглощению без существенного вклада многократного рассеяния, а концентрация поглощающей примеси относительно невелика.

Для рассматриваемой задачи контроля влажности бензина выбор закона Бугера–Ламберта–Бера в качестве базовой модели обоснован следующими соображениями. Во-первых, растворимость воды в углеводородах топливного ряда составляет, как правило, не более нескольких сотен ppm при температурах эксплуатации, что соответствует режиму молекулярного (истинного) поглощения без выраженного рассеяния. Во-вторых, длина волны источника излучения выбрана в области первого обертона валентного колебания связи O–H (порядка 1,49 мкм), где вода демонстрирует выраженную и достаточно селективную полосу поглощения на фоне относительно слабого и плавно меняющегося поглощения самого углеводородного матрикса, что также характерно для ближней инфракрасной спектроскопии топлив [6; 14]. В-третьих, конструктивно зазор между источником и приёмником в датчике мал и фиксирован, что соответствует условию однородной среды с постоянной оптической длиной пути.

Необходимо отдельно отметить, что при появлении в баке не растворённой, а свободной, эмульгированной воды — что происходит, в частности, при интенсивной механической тряске топлива во время движения автомобиля — к истинному поглощению добавляется существенный вклад рассеяния излучения на каплях воды, размеры которых сопоставимы с длиной волны излучения. Строгое описание этого случая требует перехода к теории Ми либо, при высокой кратности рассеяния, к диффузионным двух-потоковым моделям типа Кубелки–Мунка, применяемым, в частности, в конструкциях недорогих оптических датчиков мутности воды [2; 7]. В настоящей работе, ориентированной на построение первичной, инженерно-применимой модели для последующей калибровки по натурным данным, эффект рассеяния на каплях явно не выделяется в отдельное слагаемое, а неявно учитывается в составе единого эмпирически калибруемого коэффициента ослабления, что оговорено ниже при формулировке частной модели. Такое упрощение является распространённой и оправданной практикой при построении первичных калибровочных моделей недорогих оптических сенсоров [2; 7; 12] и рассматривается как первый, инженерно-достаточный уровень описания, допускающий последующее уточнение переходом к теории Ми при необходимости повышения точности.

Формализм. Рассмотрим коллимированный пучок инфракрасного излучения, распространяющийся вдоль оси, соединяющей источник излучения и фоторезистор, через слой топлива постоянной толщины. При прохождении элементарного слоя вещества относительное убывание интенсивности излучения пропорционально пройденной толщине этого слоя, а коэффициент пропорциональности определяется способностью среды поглощать излучение на данной длине волны. Это утверждение записывается в виде линейного однородного дифференциального уравнения первого порядка относительно интенсивности излучения как функции пройденного расстояния, представленного ниже в формуле (1).

$$\frac{dI(x)}{dx} = -\mu \cdot I(x) \quad (1)$$

Здесь координата x отсчитывается вдоль направления распространения пучка от источника излучения, интенсивность рассматривается как функция этой координаты, а величина μ представляет собой коэффициент ослабления среды, имеющий размерность обратной длины и зависящий от длины волны излучения, температуры и состава среды, в том числе от концентрации воды в топливе. Уравнение (1) решается разделением переменных: обе части делятся на интенсивность и интегрируются по координате от начала слоя до текущего положения. В результате разделения переменных и интегрирования обеих частей получается соотношение, связывающее натуральный логарифм отношения интенсивностей с пройденным расстоянием, приведённое в формуле (2).

$$\ln\left[\frac{I(x)}{I^0}\right] = -\mu \cdot x \tag{2}$$

В формуле (2) величина I_0 обозначает интенсивность излучения на входе в контролируемый слой топлива, то есть при нулевом значении координаты. Потенцируя обе части соотношения (2), получаем общее решение дифференциального уравнения (1), устанавливающее экспоненциальный закон убывания интенсивности излучения с расстоянием, что и составляет содержание закона Бугера–Ламберта–Бера в его классической форме, приведённой в формуле (3).

$$I(x) = I_0 \exp(-\mu \cdot x) \tag{3}$$

Поскольку в конструкции рассматриваемого датчика фоторезистор фиксирован на постоянном расстоянии от источника излучения, равном ширине зазора между стенками измерительной ячейки, для дальнейшего анализа удобно перейти от переменной координаты *x* к постоянному конструктивному параметру — длине оптического пути *L*, представляющей собой ширину зазора между источником и приёмником излучения. Соответствующая форма закона, связывающая измеряемое на выходе значение интенсивности с шириной зазора, приведена в формуле (4).

$$I(L) = I_0 \exp(-\mu \cdot L) \tag{4}$$

Отношение выходной интенсивности к входной принято называть пропусканием среды; данная величина непосредственно связана с относительным изменением фотопроводимости используемого фоторезистора и, соответственно, является измеряемым на выходе датчика сигналом. Поскольку целью работы является не только качественное, но и количественное описание зависимости показаний датчика от влажности топлива, коэффициент ослабления μ должен быть представлен как явная функция концентрации воды в бензине. В первом, простейшем приближении, справедливом в относительно узком рабочем диапазоне концентраций, данная зависимость принимается линейной, что отражено в формуле (5).

$$\mu(c) = \mu_0 + k \cdot c \tag{5}$$

В формуле (5) величина *c* обозначает концентрацию воды в бензине, параметр μ_0 представляет собой базовый коэффициент ослабления практически безводного топлива, обусловленный собственным поглощением углеводородной матрицы, оптическими потерями на границах раздела сред и конструктивными факторами измерительной ячейки, а параметр *k* характеризует чувствительность датчика к присутствию воды и объединяет в себе как истинное селективное поглощение растворённой влаги, так и — в диапазоне концентраций, где появляется свободная фаза, — дополнительный вклад рассеяния излучения на эмульгированных каплях воды, оговорённый в разделе «Материалы и методы». Подстановка выражения (5) в формулу (4) даёт итоговую общую форму математической модели датчика, связывающую измеряемое пропускание одновременно с шириной зазора измерительной ячейки и концентрацией воды в топливе, приведённую в формуле (6).

$$I(L,\ c) = I_0 \exp[-(\mu_0 + k \cdot c) \cdot L] \quad (6)$$

Формула (6) представляет собой общий вид математической модели датчика и содержит два неизвестных калибровочных параметра — μ_0 и k, — определение которых на основании граничных (калибровочных) условий рассматривается в следующем разделе.

Обсуждение

Для перехода от общей формы модели (6) к её частной, пригодной для практического использования форме необходимо определить численные значения двух калибровочных параметров — базового коэффициента ослабления *μ0* и чувствительности датчика к влаге *k*. Поскольку модель (6) содержит ровно два неизвестных параметра, для их однозначного определения достаточно и необходимо ровно два независимых калибровочных измерения пропускания при известных определяемых концентрациях воды в топливе.

Необходимо оговорить, что на момент подготовки настоящей статьи результаты натурных калибровочных измерений на изготовленном экспериментальном макете датчика были получены разработчиками, однако оказались недоступны авторам на момент написания модели по техническим причинам. В связи с этим ниже приводится демонстрационный расчёт на основе иллюстративных, физически правдоподобных калибровочных значений, отражающих типичный порядок величин пропускания для практически безводного и заметно обводнённого бензина; после получения окончательных экспериментальных данных численные коэффициенты модели подлежат прямому пересчёту по приведённой ниже методике без изменения структуры самой модели.

В качестве первого калибровочного условия принимается измерение пропускания практически безводного топлива с малой остаточной концентрацией воды при фиксированной ширине зазора, принятой равной базовому значению. Второе калибровочное условие соответствует измерению пропускания при заметно повышенной концентрации воды при той же ширине зазора. Оба условия совместно образуют систему двух линейных алгебраических уравнений относительно неизвестных μ_0 и *k*, приведённую в формулах (7) и (8).

$$\mu_0 + kc_1 = -\frac{1}{L_0}\ln(T_1) \quad (7)$$

$$\mu_0 + kc_2 = -\frac{1}{L_0}\ln(T_2) \quad (8)$$

В формулах (7) и (8) величины c_1 и c_2 обозначают концентрации воды в первом и втором калибровочных измерениях, T_1 и T_2 — соответствующие им измеренные значения пропускания, а L_0 — ширину зазора, при которой проводились калибровочные измерения. Вычитание уравнения (7) из уравнения (8) устраняет неизвестную μ_0 и позволяет выразить чувствительность датчика k непосредственно через разность калибровочных измерений, что приведено в формуле (9); после определения k значение μ_0 находится обратной подстановкой в любое из уравнений (7) или (8).

$$k = \frac{\ln(T_1) - \ln(T_2)}{L_0(c_2 - c_1)} \quad (9)$$

Приняв в качестве иллюстративных калибровочных значений концентрацию c_1, равную пятидесяти миллионным долям, с соответствующим пропусканием, равным 0,92, и концентрацию c_2, равную восьмистам миллионным долям, с пропусканием, равным 0,65, при базовой ширине зазора, равной пяти миллиметрам, решение системы (7)–(8) по формуле (9) даёт следующие численные значения калибровочных параметров: чувствительность датчика k составляет приблизительно девять целых двадцать шесть сотых на десять в минус пятой степени обратных миллиметров на миллионную долю концентрации, а базовый коэффициент ослабления μ_0 составляет приблизительно ноль целых двенадцать тысячных обратного миллиметра. Подстановка найденных значений в общую формулу (6) даёт частную, калиброванную форму математической модели датчика, пригодную для дальнейшего численного анализа и построения графических зависимостей.

На рисунке 1 представлена зависимость пропускания излучения от концентрации воды в бензине, построенная по калиброванной модели для четырёх значений ширины зазора, соответствующих реальному диапазону геометрических размеров горловин топливных баков — от нескольких миллиметров у наиболее узких конструкций до порядка десяти миллиметров у наиболее широких.

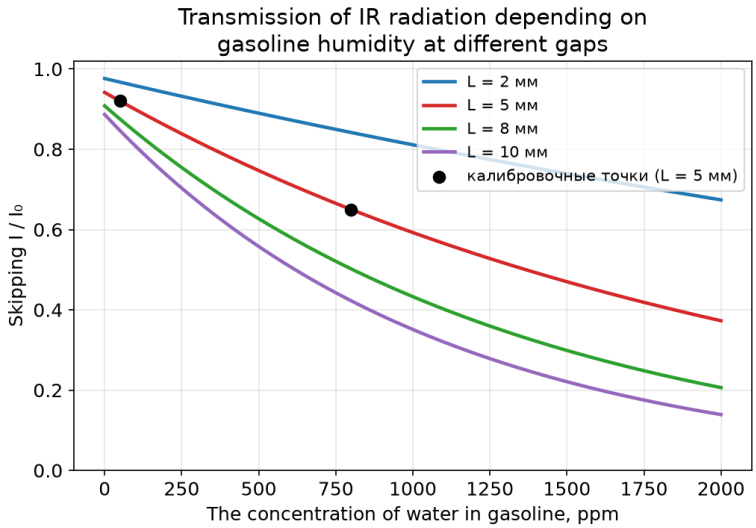

Рисунок 1. Зависимость пропускания инфракрасного излучения от концентрации воды в бензине при различной ширине измерительного зазора

Из рисунка 1 следует, что при малой ширине зазора пропускание слабо чувствительно к изменению концентрации воды и остаётся высоким практически во всём рассматриваемом диапазоне концентраций, тогда как при увеличении ширины зазора чувствительность датчика к обводнённости топлива существенно возрастает, а пропускание падает заметно быстрее уже при относительно невысоких концентрациях воды. Данное наблюдение имеет прямое конструктивное следствие: при проектировании измерительной ячейки датчика ширина зазора должна выбираться как компромисс между чувствительностью к целевому диапазону концентраций и сохранением измеримого, не слишком малого уровня выходного сигнала фоторезистора при максимальной ожидаемой обводнённости топлива.

На рисунке 2 представлена обратная по смыслу зависимость — пропускания от ширины зазора при нескольких фиксированных значениях концентрации воды, включая случай практически безводного топлива.

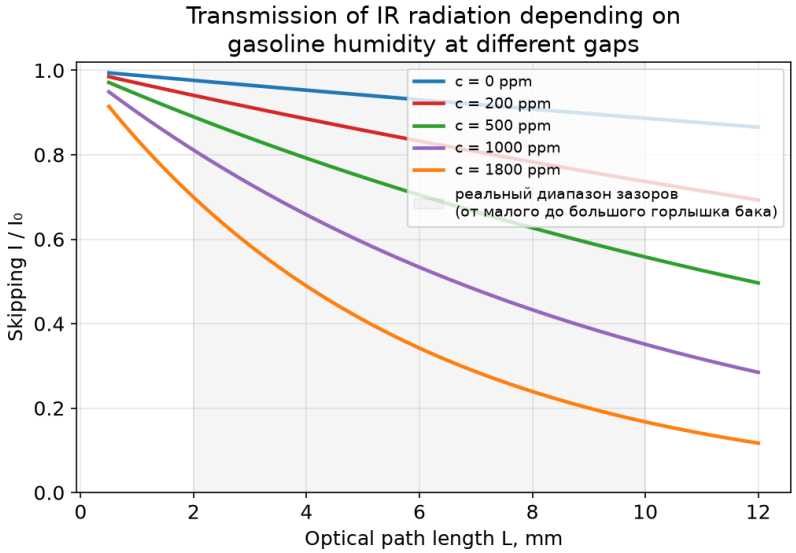

Рисунок 2. Зависимость пропускания инфракрасного излучения от ширины измерительного зазора при различной концентрации воды в бензине

Серой заливкой на рисунке выделен реальный конструктивный диапазон ширины зазора от малых до больших горловин топливных баков. Из рисунка 2 видно, что для полностью безводного топлива пропускание убывает с ростом зазора относительно медленно и остаётся на приемлемом для регистрации фоторезистором уровне во всём рассматриваемом диапазоне, тогда как при высоких концентрациях воды пропускание уже при ширине зазора порядка нескольких миллиметров падает до величин, близких к нулю, что на практике означает потерю чувствительности датчика в этом режиме и указывает на необходимость выбора достаточно малой ширины зазора именно для диапазона повышенной обводнённости, представляющего наибольший практический интерес с точки зрения предотвращения отказа двигателя.

На рисунке 3 отдельно показана калибровочная зависимость самого коэффициента ослабления от концентрации воды, полученная по формуле (5) с найденными численными значениями параметров μ_0 и k. Линейный характер данной зависимости в рамках принятого приближения является прямым следствием формулы (5) и полностью определяет нелинейный, экспоненциальный характер зависимостей пропускания, представленных на рисунках 1 и 2. Практическое значение данного графика состоит в том, что именно коэффициент ослабления, а не непосредственно измеряемое пропускание, является физически содержательной, линейно связанной с влажностью величиной, что целесообразно использовать при построении алгоритма пересчёта показаний фоторезистора в значение концентрации воды в бортовом контроллере датчика.

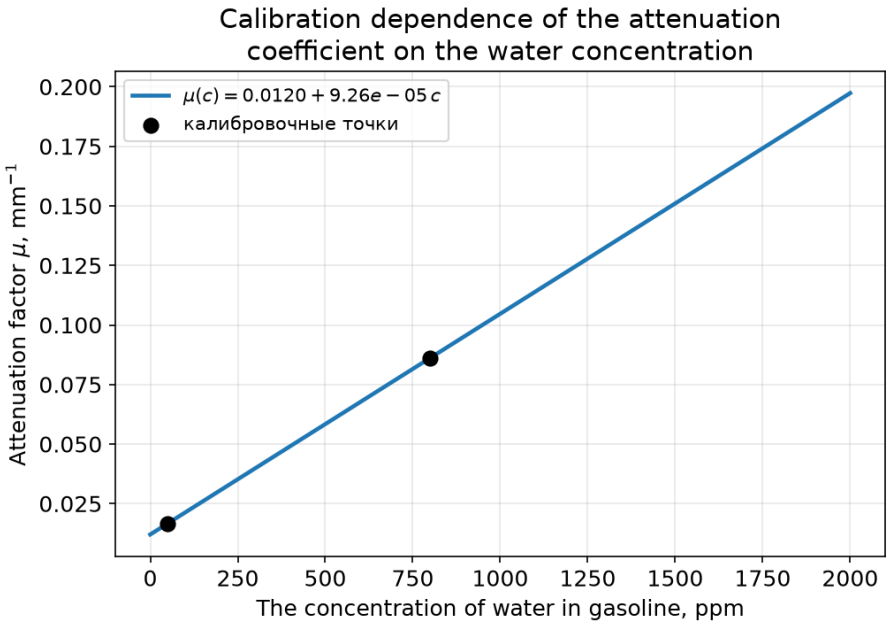

Рисунок 3. Калибровочная зависимость коэффициента ослабления от концентрации воды в бензине

Полученные результаты позволяют сформулировать два практически значимых следствия применительно к эксплуатации автомобилей специального назначения. В стационарном режиме, когда автомобиль неподвижен, а вода в баке успевает расслоиться и осесть в виде плотного нижнего слоя, датчик, установленный на определённой высоте, способен зафиксировать факт достижения уровнем воды критической отметки по резкому

падению пропускания при переходе границы раздела фаз, что позволяет заблаговременно, до выезда по экстренному вызову, произвести слив обводнённого топлива. В динамическом режиме, при движении автомобиля, когда вибрация приводит к образованию эмульсии и, соответственно, к дополнительному вкладу рассеяния в коэффициент ослабления, непрерывный мониторинг пропускания позволяет отследить нарастание обводнённости топлива по ходу движения и заблаговременно предупредить водителя о риске остановки двигателя, что представляет собой более критичный с точки зрения безопасности сценарий применения датчика по сравнению со стационарным контролем.

Заключение

В работе построена математическая модель оптического инфракрасного датчика влажности бензина, предназначенного для контроля обводнённости топлива в баках автомобилей специального назначения. Обоснован выбор закона Бугера–Ламберта–Бера в качестве базового физического приближения, получено общее аналитическое решение соответствующего дифференциального уравнения ослабления интенсивности излучения, введена линейная калибровочная зависимость коэффициента ослабления от концентрации воды и на основании модельных калибровочных данных получена частная, пригодная для практического использования форма модели. Проведённое численное моделирование позволило установить характер зависимости пропускания излучения как от концентрации воды в топливе, так и от ширины зазора измерительной ячейки датчика, а также сформулировать практические рекомендации по выбору геометрии измерительной ячейки применительно к реальному диапазону размеров горловин топливных баков.

Дальнейшее развитие представленной модели видится в двух направлениях. Во-первых, после получения окончательных экспериментальных калибровочных данных с изготовленного макета датчика необходимо провести уточнённую калибровку параметров μ_0 и k, а также оценить устойчивость модели по расширенному набору калибровочных точек, что потребует перехода от системы двух уравнений к статистической, регрессионной процедуре оценки параметров. Во-вторых, при необходимости повышения точности модели в диапазоне высоких концентраций свободной, эмульгированной воды представляется целесообразным явное выделение вклада рассеяния излучения на каплях в отдельное слагаемое коэффициента ослабления на основе теории Ми либо приближённых диффузионных моделей типа Кубелки–Мунка, что составляет предмет дальнейшего, более общего исследования.

Список литературы

1. Blau B., Krzeczek O., Heinrich C. et al. Analysis of water-in-gasoline emulsions via experiments and direct numerical simulations // Experimental and Computational Multiphase Flow. – 2026. – Т. 7. URL: https://doi.org/.
2. Bohren C.F., Huffman D.R. Absorption and Scattering of Light by Small Particles. – New York: John Wiley & Sons, 1983. – 530 P.

3. Bozdoğan A. Polynomial Equations based on Bouguer–Lambert and Beer Laws for Deviations from Linearity and Absorption Flattening // Journal of Analytical Chemistry. – 2026. – T. 77. URL: https://doi.org/.
4. Buyanov D.A., Shalaev P.V., Zabodaev S.V. et al. Measurement of Hemoglobin Oxygenation Parameters in Arterial Occlusion and Physical Exertion // Biomedical engineering. – 2026. – T. 57. URL: https://doi.org/.
5. Chandrasekhar S. Radiative Transfer. – New York: Dover Publications, 1960. – 393 P.
6. Correia R.M., Domingos E., Cáo V.M., Araujo B.R.F., Sena S., Pinheiro L.U., Fontes A.M., Aquino L.F.M., Ferreira E.C., Filgueiras P.R., Romão W. Portable near infrared spectroscopy applied to fuel quality control // Talanta. – 2018.
7. Fay C.D., Nattestad A. Advances in Optical Based Turbidity Sensing Using LED Photometry (PEDD) // Sensors. – 2022. – T. 254, № 1.
8. Gisi, M.F.S., Navratil, O., Cherqui, F. et al. From dishwasher to river: how to adapt a low-cost turbidimeter for water quality monitoring // Environmental Monitoring and Assessment. – 2026. – T. 196. URL: https://doi.org/.
9. Herzog, A., Iriawan, H., Ah, L. et al. Unveiling solid electrolyte interphase dynamics in electrochemical lithium-mediated ammonia synthesis via operando Raman spectroscopy [Электронный ресурс]. // Nature Catalysis. – 2026. URL: https://doi.org/. (дата обращения 24.09.2026)
10. Kazak, A.V., Kirillov, A.A., Simonchik, L.V. et al. Diagnosis of Bactericidal Components of Air-Plasma Jets by IR and UV Absorption Spectroscopy** // Journal of Applied Spectroscopy. – 2026. – T. 91. URL: https://doi.org/.
11. Keller N.V., Nikolkin V.N., Butakov D.S. et al. Optimization of the Technology for Manufacturing the Electrodes for Self-Charging Supercapacitors from Carbon Nanotubes // Russian Journal of Electrochemistry. – 2026. – T. 60. URL: https://doi.org/.
12. Kiteto M.K., Mecha, C.A. A unified electromagnetic framework for light absorption: beyond the Bouguer–Beer–Lambert law // Discover Chemistry. – 2026. – T. 2. URL: https://doi.org/.
13. Masiovskij Ł., Sobczyński D. Macioszek Ł., Sobczyński D. Moisture Content Assessment of Commercially Available Diesel Fuel Using Impedance Spectroscopy // Energies. – 1903. – T. 1903, № 8.
14. Rocher J., Jimenez J.M., Tomas J., Lloret J. Low-Cost Turbidity Sensor to Determine Eutrophication in Water Bodies // Sensors. – 2023. – T. 3913, № 8.
15. Sharma A., Mishra S., Vishwakarma P.K. et al. Visibility analysis of diffusion flames in methanol–gasoline blended pool fires // Journal of the Brazilian Society of Mechanical Sciences and Engineering. – 2026. – T. 47. URL: https://doi.org/.
16. Sinimbu L.I.M., Carvalho, T.C.V., De Falco, A. et al. Using the Beer–Lambert law to determine the Euler number using a Markovian system // Indian Journal of Physics. – 2026. – T. 98. URL: https://doi.org/.

17. Zevatskii Y.E., Lysova S.S., Skripnikova T.A. et al. Bouguer–Lambert–Beer Law of Absorption: Spectrophotometry in Electrolyte Solutions // Russian Journal of Physical Chemistry. – 2026. – T. 98. URL: https://doi.org/.

ДЛЯ ЗАМЕТОК

ДЛЯ ЗАМЕТОК

UNIVERSUM: ТЕХНИЧЕСКИЕ НАУКИ

Электронный научный журнал

№ 9(150)
Сентябрь 2026 г.
Часть 1

Свидетельство о регистрации СМИ: ЭЛ № ФС 77 – 54434 от 17.06.2013

Издание содержит научную информацию и в соответствии с ч. 4 ст. 1 Федерального закона от 29.12.2010 № 436-ФЗ «О защите детей от информации, причиняющей вред их здоровью и развитию» не подлежит классификации (маркировке) по возрастным категориям

Издательство ООО «Юниверсум»

Объём издания: 88 с.

Минимальные системные требования: процессор с частотой от 1,3 ГГц; ОЗУ от 512 МБ; программа для чтения PDF-файлов; доступ к сети Интернет

Дата размещения в сети Интернет: 28.09.2026

Журнал размещается в:

CYBERLENINKA

Google scholar

BASE

OpenAIRE

НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА eLIBRARY.RU

ULRICHSWEB GLOBAL SERIALS DIRECTORY

СОЦИОНЕТ

EBSCO

Registry of Open Access Repositories (ROAR)