
Компьютерные инструменты в образовании, 2024
№ 1: 32–44
УДК: 004.8:378.145
http://cte.eltech.ru
doi:10.32603/2071-2340-2024-1-32-44

# ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ ПРОМПТ-ИНЖИНИРИНГА И КВАНТОВАННЫХ LLM В СОЗДАНИИ СТРУКТУРЫ АКАДЕМИЧЕСКИХ КУРСОВ

Шнайдер П. А.[1], аспирант, ассистент, beatrix.linkoln@gmail.com, orcid.org/0000-0002-4147-6561
Чернышева А. В.[1], ассистент, ✉ avchernysheva@itmo.ru, orcid.org/0000-0002-9956-6607
Никифорова А. Д.[1], студент, 34743@niuitmo.ru
Говоров А. И.[1], старший преподаватель, govorov@itmo.ru, orcid.org/0009-0005-6674-1666
Хлопотов М. В.[1], канд. техн. наук, доцент, khlopotov@itmo.ru, orcid.org/0000-0002-9053-027X

[1]Национальный исследовательский университет ИТМО,
Кронверкский пр., 49, лит. А, 197101, Санкт-Петербург, Россия

**Аннотация**

В данной статье представлены итоги эксперимента по применению больших языковых моделей (LLM) для создания структуры университетских курсов. Для формирования запросов к LLM использовались такие методы промпт-инжиниринга, как zero-shot, few-shot, chain-of-thought и tree-of-thought. Для эксперимента преимущественно использовались квантованные модели, такие как mistral-7b-instruct, mixtral-8x7b-instruct, openchat_3.5, saiga2_13b, starling-lm-7b-alpha, tinyllama и другие. Сгенерированные ими структуры курсов сравнивались с данными, полученными с помощью ChatGPT-4. Модели openchat_3.5.q5_k_m и starling-lm-7b-alpha.q5_k_m показали сопоставимое с ChatGPT-4 качество генерации рабочих программ дисциплин. Эксперимент подчеркивает возможности применения LLM в сфере образования и указывает на перспективные направления для дальнейших исследований.

**Ключевые слова:** *большие языковые модели, промпт-инжиниринг, квантованные модели, few-shot, zero-shot, chain-of-thought*

**Цитирование:** Шнайдер П. А., Чернышева А. В., Никифорова А. Д., Говоров А. И., Хлопотов М. В. Исследование эффективности промпт-инжиниринга и квантованных LLM в создании структуры академических курсов // Компьютерные инструменты в образовании. 2024. № 1. С. 32-44. doi:10.32603/2071-2340-2024-1-32-44

## 1. ВВЕДЕНИЕ

Большая языковая модель (LLM, Large Language Model) — это тип модели глубокого обучения, которая понимает и генерирует текст на естественном языке. Эти модели обучаются на огромных объемах текстовых данных и содержат в себе большое число параметров, что позволяет им распознавать, переводить, прогнозировать и генерировать текст или другой контент [1].

Благодаря этим способностям большие языковые модели находят применение в множестве областей, включая образование. В образовательной сфере их применение открывает новые возможности как для преподавателей, так и для студентов, включая персонализированное обучение и поддержку обучения на другом языке [2].

Применение языковых моделей в образовании также может помочь сделать обучение более доступным для людей с ограниченными возможностями, предлагая альтернативные способы обучения и коммуникации [3]. В области образования LLM могут служить инструментом для проведения исследований, подготовки материалов к занятиям и первичной проверки домашних заданий [4].

В данной работе рассматривается применение LLM для генерации структуры университетских курсов. Предположительно, структура курса, сгенерированная с помощью LLM, поможет преподавателям создавать курсы, учитывающие конкретные требования учебного заведения и образовательной программы [5].

Так, например, в курсе по нейронным сетям после изучения основ могут следовать более сложные темы, такие как глубокое обучение и специализированные нейронные сети, а затем этические аспекты их использования [6].

Структура курса должна учитывать актуальные тенденции в предметной области и сохранять баланс между широтой тематического охвата и углубленным изучением отдельных тем [7].

## 2. ПРОМПТ-ИНЖИНИРИНГ

В качестве инструментального средства воздействия на большую языковую модель использованы методики промпт-инжиниринга. Промпт-инжиниринг — это процесс оптимизации запросов (промптов) для взаимодействия с искусственными интеллектуальными системами, особенно с большими языковыми моделями.

Целью промпт-инжиниринга является формирование таких запросов, которые наиболее эффективно и точно направляют алгоритмы искусственного интеллекта (ИИ) к требуемым результатам. Это включает в себя подбор определенных слов, фраз и контекстуальных подсказок, которые могут улучшить понимание запроса ИИ и увеличить вероятность получения релевантного и точного ответа.

В проводимом эксперименте были задействованы четыре техники промпт-инжиниринга: zero-shot [8], few-shot [9], chain-of-thought [10], tree-of-thought [11]. Далее представлено краткое описание каждой из них.

В подходе zero-shot большая языковая модель получает задачу без дополнительных примеров и пытается ее решить, используя свой предыдущий опыт. Примером промпта может послужить: *«Опиши структуру курса по Web-программированию»*, без предварительных примеров или объяснений.

Во few-shot модели предоставляется несколько примеров. Это помогает модели понять контекст и желаемый формат ответа. Данный подход эффективен, когда требуется направить модель на определенный стиль ответа или когда модель должна понять конкретный контекст задачи.

Ниже приведен пример запроса к LLM для генерации программы курса по веб-разработке с элементами few-shot. Для примеров использованы темы ранее разработанных аналогичных программ дисциплин в университете. Исходя из этих примеров LLM должна понять ожидаемую степень детализации тем в курсе, а также соотношение углубленности и обзорности материала.

*«Разработай структуру курса по дисциплине “Web-программирование” для студентов бакалавриата с акцентом на практическую сторону. Курс должен включать изучение фреймворков Django и Vue.js и охватывать темы, связанные с работой с сокетами, сетевой моделью OSI, Apache vs Nginx, ООП и паттернами проектирования, включая MVC. Для лучшего понимания структуры включи в запрос следующие примеры:*

*Пример 1:*
*Основы Web-программирования*
*Введение в HTML, CSS и JavaScript*
*Структура веб-страницы, Основы стилизации, Основы программирования на JavaScript*

*Пример 2:*
*Продвинутое Web-программирование*
*Использование фреймворков Django и Vue.js*
*Архитектура Django, Основы Vue.js, Связывание данных и компонентный подход*

*Пример 3:*
*Разработка информационной системы и интеграция пользовательского интерфейса*
*Шаблонизация пользовательских представлений*
*Использование шаблонов в Django, Взаимодействие с шаблонами в Vue.js»*

Подход chain-of-thought стимулирует модель последовательно разъяснять свой ход мыслей. Эта техника особенно ценна при решении задач, требующих глубокого логического анализа как в математике, так и при работе с абстрактными понятиями. Применительно к разработке учебных программ, данный метод способствует формированию логических связей между разделами и темами курса. Также он обеспечивает структурированное усвоение материала студентами, гарантируя, что каждый последующий модуль строится на знаниях, полученных в предыдущих разделах. Ниже приведен пример запроса в данной технике:

*Напиши структуру курса по дисциплине «Web-программирование», используя подход chain-of-thought. Начни с анализа того, какие знания, умения и навыки у студентов уже есть до начала курса. Затем для каждого раздела курса объясни, почему ты выбрал именно эти темы и подтемы и как они связаны с предыдущими знаниями студентов. Включи размышления о том, как каждая тема подготавливает студентов к последующим темам, обоснуй, почему ты считаешь, что после изучения одной темы студенты готовы перейти к следующей. Приведи примеры заданий или проектов, которые помогут закрепить полученные знания и навыки. Подведи итог, объясняя, как вся структура курса способствует достижению образовательных целей дисциплины «Web-программирование».*

Подход tree-of-thought в определенной степени можно рассматривать как расширение концепции chain-of-thought. Здесь модель одновременно рассматривает несколько путей рассуждения, что похоже на структуру дерева. Данный метод применяется для решения задач высокой сложности, которые подразумевают многочисленные возможные решения или требуют всестороннего анализа различных аспектов задачи. Ниже приводится пример запроса, включающего элементы tree-of-thought. Здесь модель должна имитировать конкурентную дискуссию, в результате которой необходимо достичь консенсуса в плане тем, включаемых в структуру курса.

*Смоделируй следующую ситуацию: 100 экспертов составляют курс по дисциплине «Web-программирование». Каждый должен включить в свой курс 5–7 тем. По итогу получается 100 различных рабочих программ дисциплин. Твоя задача — на основании*

*этих рабочих программ составить список обязательных к освоению тем, исключив из него темы, встречающиеся менее чем в 5 рабочих программах экспертов. В качестве ответа приведи общий список тем.*

## 3. ОПИСАНИЕ ЭКСПЕРИМЕНТА

Был проведен эксперимент по генерации 12 программ академических курсов из разных предметных областей, а именно:

1. Компьютерные сети.
2. Управление информационной безопасностью.
3. Иностранный язык в профессиональной деятельности.
4. Веб-программирование.
5. Основы экономики.
6. Технологический форсайт.
7. Математическая лингвистика.
8. Анализ и проектирование на UML.
9. Гибкий менеджмент.
10. Физика лазеров.
11. Искусство, наука и технологии.
12. Биометрия и нейротехнологии.

Для каждого курса было сформулировано по четыре запроса с использованием вышеупомянутых техник промпт-инжиниринга. Для few-shot использовались примеры из существующих программ учебных курсов, разработанных преподавателями университета. Полученный набор промптов был отправлен на вход десяти LLM. Ввиду ограниченности вычислительных ресурсов девять из них были квантованными.

В контексте большой языковой модели квантование — это процесс оптимизации, при котором веса модели преобразуются в представление с более низкой точностью. Например, используются 8-битные целые числа вместо 32 или 16-битных с плавающей точкой. Это ведет к снижению качества ответов LLM, однако во многих случаях снижение точности несущественно по сравнению с преимуществами сокращения размера и увеличения скорости [12, 13].

На текущий момент доступно множество квантованных больших языковых моделей (LLM), показывающих, несмотря на свой небольшой размер, впечатляющие результаты. Для эксперимента была выбрана легковесная модель Mixtral 8X7B Instruct от Mistral AI с архитектурой mixture of experts [14], использующая двухуровневый метод квантования весов.

В исследовании также использовалась модель OpenChat 3.5 7B. Эта модель выделяется своей выдающейся производительностью [15] в обработке естественного языка. В ней используется метод квантования *Q5_K_M*, обеспечивающий эффективное 5-битное квантование. Несмотря на сравнительно небольшое количество параметров для LLM — всего 7 миллиардов — и размер в 5.13 ГБ, модель требует до 7.63 ГБ оперативной памяти для работы. Благодаря данным характеристикам, результаты, достигаемые моделью *OpenChat 3.5 7B*, сравнимы с результатами *ChatGPT*, что делает её привлекательной для использования в проектах, где важно минимизировать потерю качества.

Starling LM 7B Alpha, разработанная командой Berkeley-Nest, представляет собой языковую модель, дообученную с использованием методов обучения с подкреплением

(RLHF/RLAIF) на основе модели Openchat 3.5. Эта модель обладает высокой производительностью благодаря применению нового метода оптимизации политик, известного как Advantage-Induced Policy Alignment (APA), и использованию обучающего набора данных berkeley-nest/Nectar. С помощью метода квантования Q5_K_M и 5-битной точности, модель занимает 5.13 ГБ дискового пространства и требует до 7.63 ГБ оперативной памяти, что делает ее подходящей для крупномасштабных задач с минимальной потерей качества. Starling LM 7B Alpha показала [16] впечатляющие результаты, достигнув оценки 8.09 в MT Bench с GPT-4 в качестве судьи, превзойдя все модели на MT-Bench, кроме *GPT-4* и *GPT-4 Turbo* от OpenAI.

Дополнительно были взяты еще несколько квантованных моделей. Ниже приведен итоговый список:

1. *mistral-7b-instruct-v0.1.Q5_K_M*.
2. *mixtral-8x7b-instruct-v0.1.Q2_K*.
3. *openbuddy-mistral-7b-v13.Q5_K_M*.
4. *openbuddy-llama2-13b64k-v15.Q5_K_M*.
5. *openchat_3.5.Q5_K_M*.
6. *saiga2_13b*.
7. *starling-lm-7b-alpha.Q5_K_M*.
8. *synthia-7b-v1.3.Q4_K_M*.
9. *tinyllama-1.1b-1t-openorca.Q5_K_M*.

Кроме того, проводился эксперимент с использованием ChatGPT-4. Эта модель обладает значительно большим числом параметров по сравнению с упомянутыми ранее и предъявляет более высокие требования к аппаратным ресурсам. Отсюда разумно предположить, что ChatGPT-4 должен показывать лучшую производительность. Однако интеграция ChatGPT в сервис не всегда возможна. Во многих случаях более доступные и менее требовательные к ресурсам модели демонстрируют сопоставимые результаты для специфических задач, особенно при эффективном использовании техник промптинга, как это было показано в исследовании [17].

## 4. ОЦЕНКА РЕЗУЛЬТАТОВ

В ходе вычислений были получены 480 текстовых фрагментов, описывающих структуру курса: 12 дисциплин × 4 подхода к промптингу × 10 больших языковых моделей. Помимо экспертной оценки, для определения качества генерации LLM использовался алгоритм формирования эмбеддинга.

Построение эмбеддинга дисциплины происходит через граф учебных сущностей. Под учебными сущностями понимается множество пререквизитов и результатов изучения дисциплины — навыки, которыми обладает студент до и после прохождения курса. На графе между двумя вершинами есть связь, если две соответствующие им сущности встретились в описании одной дисциплины.

Далее в несколько этапов формируется эмбеддинг:

1. Создание графа предметных областей $G_d(E, V)$ путем кластеризации графа учебных сущностей. Предметной областью считается кластер, в который входит более 10 учебных сущностей.
2. Распределение учебных сущностей из небольших кластеров (до 10 сущностей) по сформированным предметным областям на основе контекстной близости токенов

учебной сущности с токенами из предметной области. Контекстная близость вычисляется на основе эмбеддингов модели Word2Vec, обученной на данных 8699 дисциплин, реализованных в Университете ИТМО в 2018–2023 гг.

3. Подсчет количества сообществ $n$ на получившемся графе предметных областей $G_d(E, V)$ . Формирование нулевого вектора размером $n$. Для данных в описываемом эксперименте размер вектора $n = 95$.
4. Вычисление количества учебных сущностей из предметной области $D_j$, $j = \overline{1, ..., n}$, входящих в описание каждой дисциплины $C_i$, $i = \overline{1, ..., m}$. В таблице 1 приведен общий вид набора эмбеддингов размером для дисциплин.
5. Нормализация эмбеддинга дисциплины.

**Таблица 1.** Пример формирования эмбеддинга

| | $D_0$ | $D_1$ | $D_2$ | … | $D_n$ |
|---|---|---|---|---|---|
| $C_0$ | 2 | 6 | 0 | … | 3 |
| … | … | … | … | … | … |
| $C_m$ | 1 | 9 | 0 | … | 0 |

Такой подход к построению эмбеддингов позволяет зафиксировать и оценить общую предметную направленность курса, не углубляясь в детали терминологии, что особенно важно при сравнении объемных текстов с большой вероятностью употребления разных понятий для описания близких концепций.

Сформированный по сгенерированному тексту эмбеддинг сравнивался с эмбеддингом существующей дисциплины через косинусное сходство: чем ближе к единице, тем более похожими являются векторы.

На рисунках 1 и 2 представлено среднее сходство эмбеддингов с эталонной дисциплиной. Как видно из рисунка 1, лучший результат получился у LLM *chatgpt-4*, *openchat_3.5* и *starling-lm-7b-alpha*.

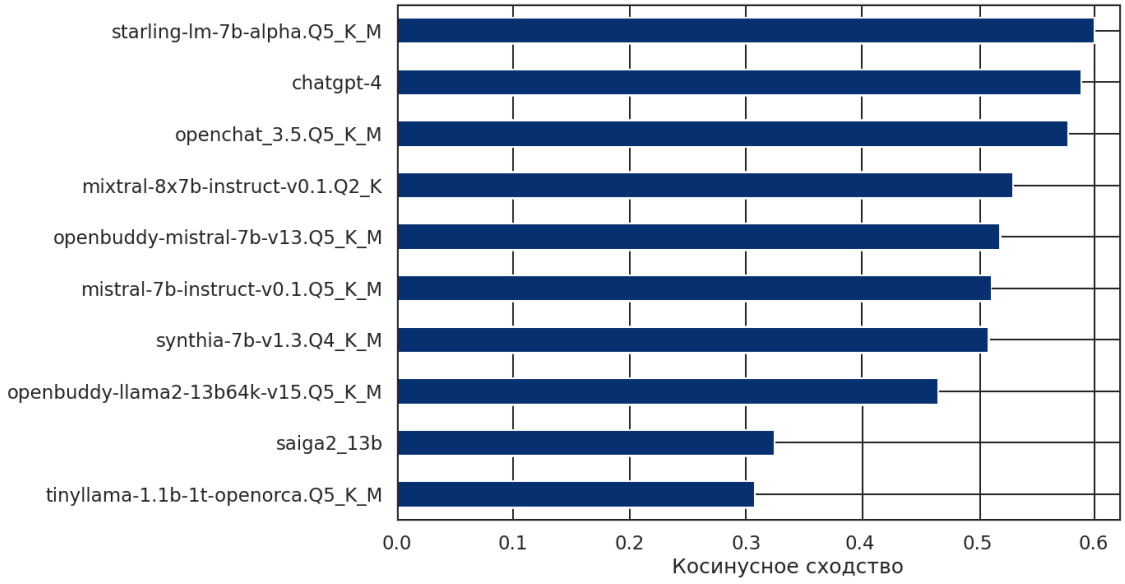


**Рис. 1.** Среднее сходство эмбеддингов сгенерированного контента с эталоном для каждой модели

На рисунке 2 представлена статистика среднего сходства эмбеддинга с эталоном для каждого метода промптинга. Лучший результат показал few-shot подход, на втором месте находится chain-of-thought, худшие результаты были получены для tree-of-thoughts.

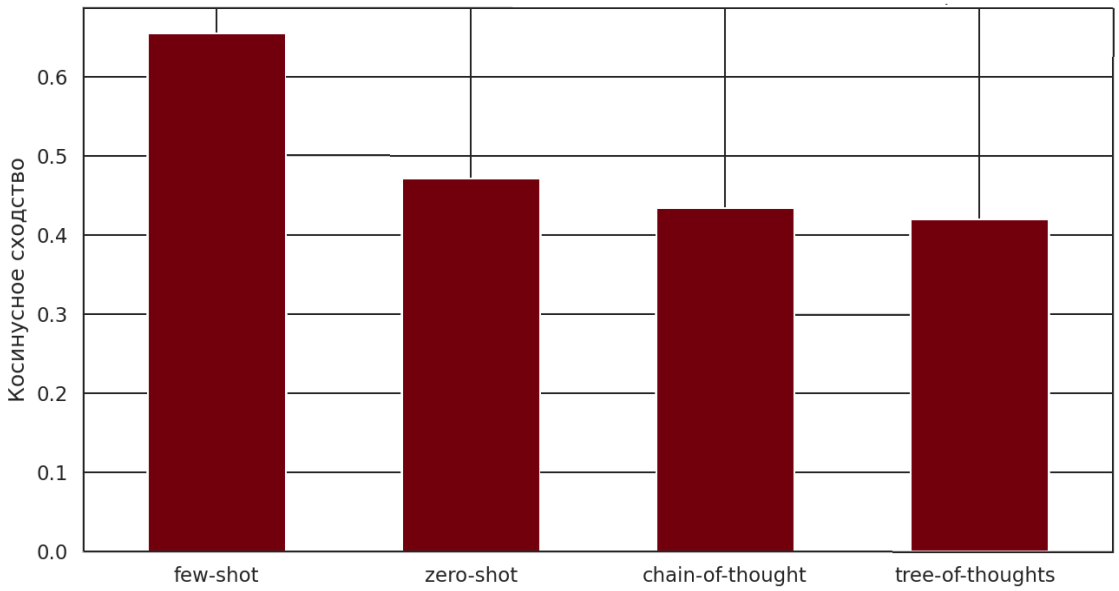


**Рис. 2.** Среднее сходство эмбеддингов сгенерированного контента с эталоном для каждого метода промптинга

На рисунке 3 представлены результаты, полученные каждой квантованной моделью и каждым методом промптинга без усреднения. Числовые значения в матрицах являются косинусным сходством сгенерированного содержания дисциплины с эталоном, составленным автором курса. Видна зависимость успешности предсказаний от используемой техники промптинга и отдельной LLM. Особенно выделяются модели *openchat_3.5.Q5_K_M*, *starling-lm-7b-alpha.Q5_K_M* и *openbuddy-mistral-7b-v13.Q5_K_M*, которые показывают высокую эффективность при любом выбранном подходе промптинга. В то же время LLM *tinyllama-1.1b-1t-openorca.Q5_K_M* демонстрирует низкие результаты в технике chain-of-thought, а модель *saiga2_13b* показывает низкую эффективность при использовании техник chain-of-thought и few-shot.

Полученный результат может быть объяснен следующими факторами: во-первых, размер модели и количество ее параметров влияют на способность к глубокому пониманию контекста и логическому рассуждению, что особенно важно для chain-of-thought. Во-вторых, эффективность модели сильно зависит от её обучающих данных и методов файнтюнинга. Важно отметить, что в ходе эксперимента промпты не адаптировались под конкретную LLM, за исключением требуемого шаблона. Вероятно, при более тщательной формулировке запросов могли быть достигнуты лучшие результаты. Производительность модели может снижаться, если она не была специально дообучена для выполнения задач, требующих сложных рассуждений. В-третьих, специализация модели играет важную роль. Saiga, например, может быть оптимизирована для конкретных целей, таких как понимание естественного языка, генерация текста, перевод, суммаризация, ответы на вопросы или диалоговые системы. В зависимости от её уникальной настройки и обучения [18] Saiga может демонстрировать различные уровни эффективности в применении техник chain-of-thought и few-shot.

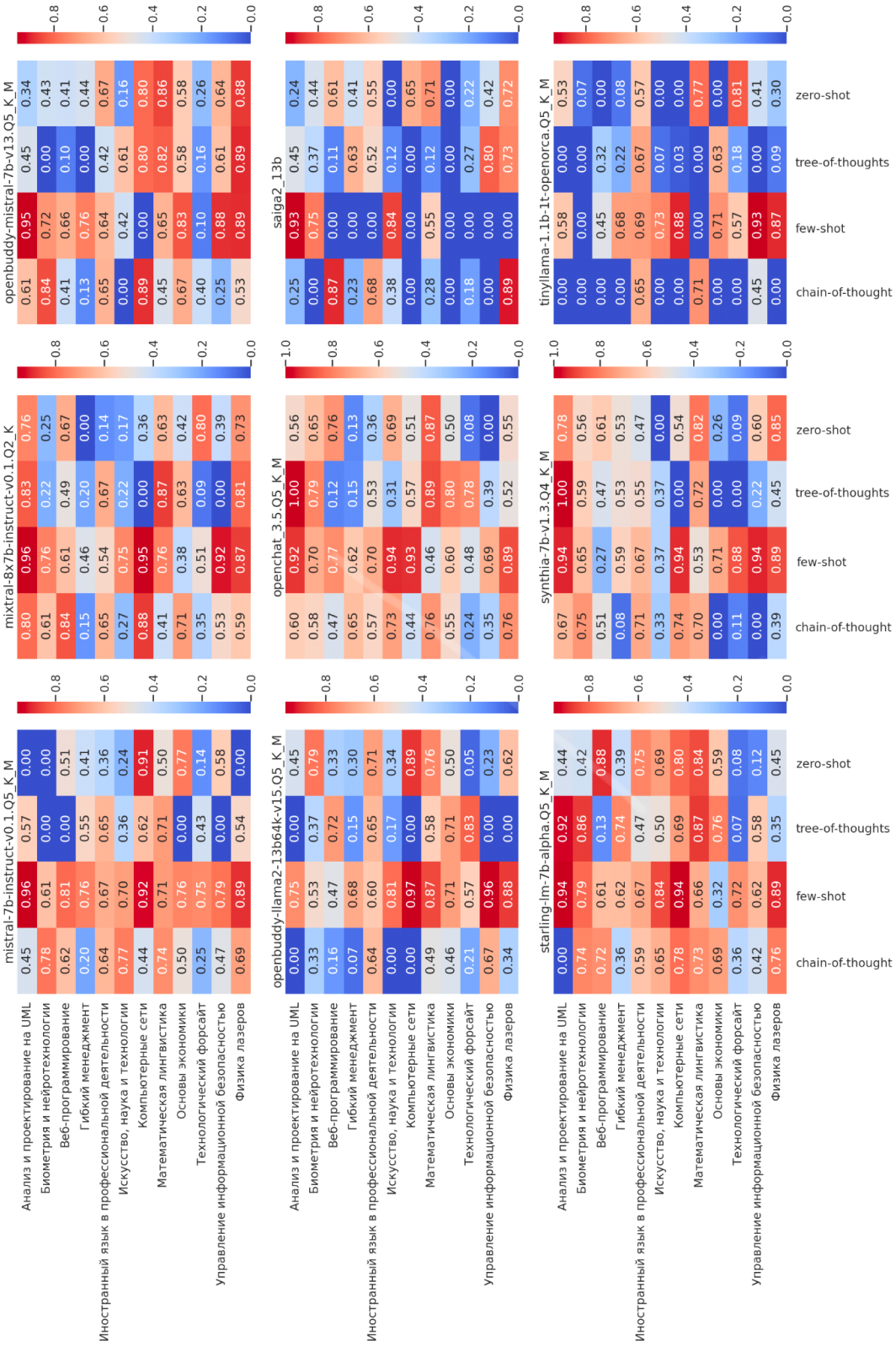


**Рис. 3.** Диаграмма разницы эмбеддингов с эталоном без усреднения для квантованных моделей

Также сходство эмбеддингов нельзя назвать единственной верной оценкой полученной структуры курса. Если сгенерированный набор тем мало пересекается с темами существующего курса, это не значит, что он некорректный. Дисциплины с одним и тем же названием могут освещать различный набор аспектов, и разработка академического курса — в большей степени творческая задача.

LLM ChatGPT показала наилучшие результаты (рис. 4) по сравнению с другими моделями, особенно в технике few-shot. Однако тут важно отметить два фактора: во-первых, ChatGPT-4 на текущий момент является одной из самых сложных и требовательных по ресурсам моделей, и некорректно сравнивать ее с менее ресурсозатратными альтернативами. Во-вторых, как уже было замечено, для few-shot использовались примеры из уже существующих дисциплин, которые и являлись эталонами при сравнении эмбеддингов. С одной стороны, это означает, что подход few-shot хорошо работает в задаче генерации структуры курса. С другой — пока неясно, насколько хорошо справится LLM с более творческой задачей, где примеры разделов и тем будут взяты не из эталона.

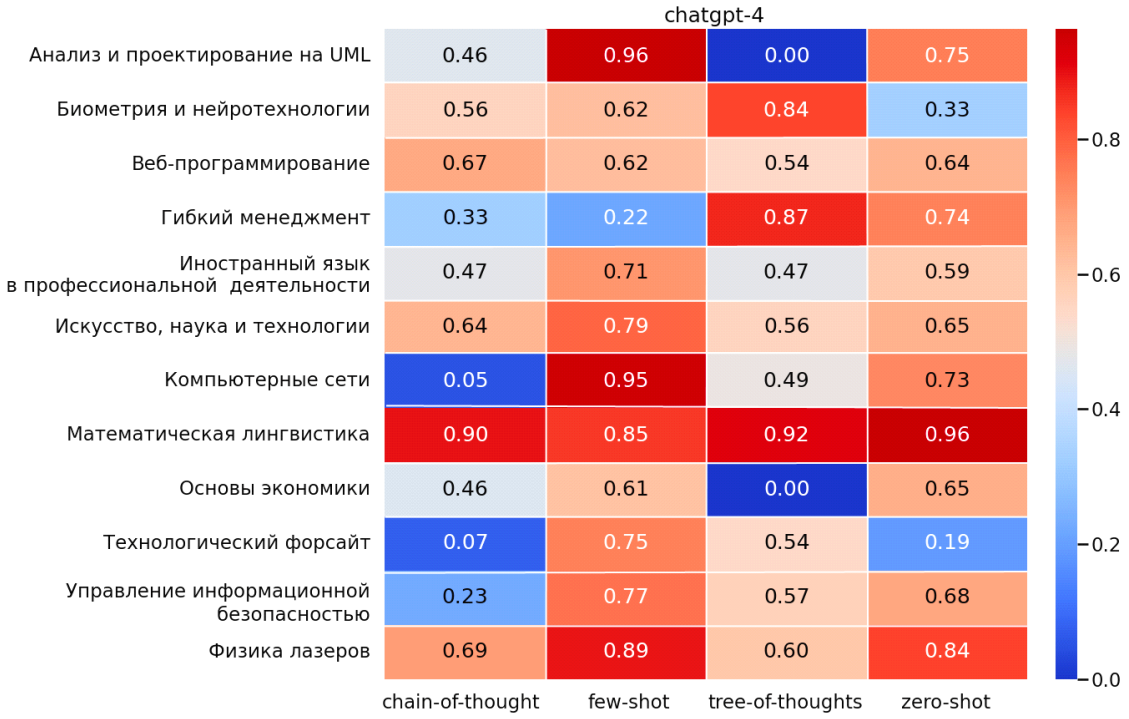


**Рис. 4.** Диаграмма сходства эмбеддингов с эталоном без усреднения для ChatGPT-4

## 5. ЗАКЛЮЧЕНИЕ И ДАЛЬНЕЙШАЯ РАБОТА

Целью проведенного эксперимента в первую очередь было получение общего представления об эффекте, который оказывает промпт-инжиниринга на процесс формирования структуры учебных дисциплин. Такие LLM как *synthia-7b-v1.3.Q4_K_M*, *openchat_3.5.Q5_K_M*, *openbuddy-mistral-7b-v13.Q5_K_M* и *starling-lm-7b-alpha.Q5_K_M* выглядят хорошими кандидатами для дальнейшего исследования.

Также необходимо выяснить, насколько успешно модели будут работать с промтами в технике few-shot, сформированными без использования примеров из эталонных курсов, и способна ли модель на творчество.

В дополнение к этому предполагается внедрение подхода Retrieval-Augmented Generation (RAG) [19] для усовершенствования процесса создания контента по дисциплине. Данный подход дополняет внутренние данные актуальной информацией из внешних источников, что способствует повышению качества и точности сгенерированных ответов, особенно в контексте сложных и специализированных тем, таких как разработка образовательного контента.

Также планируется разработать дополнительные метрики оценки качества результата, помимо косинусного сходства эмбеддингов.

## Список литературы


1. *Jermakowicz E. K.* The Coming Transformative Impact of Large Language Models and Artificial Intelligence on Global Business and Education. Journal of Global Awareness, 4(2), Article 3. (2023).
2. "AI-Assisted Learning with ChatGPT and Large Language Models: Implications for Higher Education"(2023). The 23rd IEEE International Conference on Advanced Learning Technologies, doi: 10.1109/ICALT58122.2023.00072, Orem, Utah, United States. P. 226–230.
3. *Giretti A., Durmus D., Vaccarini M., Zambelli M., Guidi A., Meana F.* INTEGRATING LARGE LANGUAGE MODELS IN ART AND DESIGN EDUCATION. 2023. P. 1–7.
4. *Jeon J., Lee S.* Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. Education and Information Technologies. 2023. Vol. 28, № 12. P. 15873–15892. doi:10.1007/s10639-023-11834-1.
5. *Abedi M., Alshybani I., Shahadat M. R. B., Murillo M.* Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education. 2023. doi:10.32388/MD04B0.2.
6. *Prather J., Denny P., Leinonen J., Becker B., Albluwi I., Craig M., Keuning H., Kiesler N., Kohn T., Luxton-Reilly A., Macneil S., Petersen A., Pettit R., Reeves B., Savelk, J.* The Robots Are Here: Navigating the Generative AI Revolution in Computing Education. 2023. doi:10.1145/3623762.3633499.
7. *Irfan M., Murray L.* Micro-Credential: A Guide to Prompt writing and Engineering in Higher Education: A tool for Artificial Intelligence in LLM. 2023. doi:10.13140/RG.2.2.15596.95367.
8. *Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa.* Large Language Models Are Zero-Shot Reasoners. Advances in Neural Information Processing Systems 35 2022. P. 22199–22213.
9. *Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei.* Language Models Are Few-Shot Learners. In Advances in Neural Information Processing Systems, 2020. Vol. 33. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
10. *Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models // In Advances in Neural Information Processing Systems, 2022. Vol. 35. P. 1–14. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
11. *Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K.* Tree of Thoughts: Deliberate Problem Solving with Large Language Models. 2023. https://arxiv.org/abs/2305.10601
12. *Li S., Ning X., Ke H., Liu T., Wang L., Li X., Zhong K., Dai G., Yang H., Wang Y.* LLM-MQ: Mixed-precision Quantization for Efficient LLM Deployment. 2023. P. 1–5.
13. *Li W., Hu A., Xu N., He G.* Quantization and Hardware Architecture Co-Design for Matrix-Vector Multiplications of Large Language Models. IEEE Transactions on Circuits and Systems I: Regular Papers, PP, 2024. P. 1-4. doi:10.1109/TCSI.2024.3350661.
14. *Jiang A. Q., Sablayrolles A., Roux A., Mensch A., Savary B., Bamford C., El Sayed W.* Mixtral of Experts. 2024. doi:10.48550/arXiv.2401.04088

15. *Wang G., Cheng S., Zhan X., Li X., Song S., Liu Y.* OpenChat: Advancing Open-source Language Models with Mixed-Quality Data. arXiv:2309.11235 [cs.CL]. 2023. doi:10.48550/arXiv.2309.11235
16. *Zhu B., Frick E., Wu T., Zhu H., Jiao J.* Starling-7B: Improving LLM Helpfulness & Harmlessness with RLAIF. По состоянию на 28.01.2024, 2023. доступно на https://starling.cs.berkeley.edu/
17. *Irugalbandara C., Mahendra A., Daynauth R., Kasthuri Arachchige T., Flautner K., Tang L., Kang Y., Mars J.* A Trade-off Analysis of Replacing Proprietary LLMs with Open Source SLMs in Production. arXiv:2312.14972 [cs.SE]. 2024. doi:10.48550/arXiv.2312.14972
18. *Tikhomirov M., Chernyshev D.* Impact of Tokenization on LLaMa Russian Adaptation. arXiv:2312.02598 [cs.CL]. 2023. doi:10.48550/arXiv.2312.02598
19. *Chen J., Lin H., Han X., Sun L.* Benchmarking Large Language Models in Retrieval-Augmented Generation. arXiv:2309.01431 [cs.CL] 2024. doi:10.48550/arXiv.2309.01431

*Поступила в редакцию 29.01.2024, окончательный вариант — 29.02.2024.*

**Шнайдер Полина Анатольевна, аспирант 3-го года, ассистент, факультет инфокоммуникационных технологий, ИТМО, beatrix.linkoln@gmail.com**

**Чернышева Анастасия Вадимовна, ассистент, факультет инфокоммуникационных технологий, ИТМО, ✉ avchernysheva@itmo.ru**

**Никифорова Анна Дмитриевна, студент 3-го курса бакалавриата, факультет инфокоммуникационных технологий, ИТМО, 34743@niuitmo.ru**

**Говоров Антон Игоревич, старший преподаватель, факультет инфокоммуникационных технологий, ИТМО, govorov@itmo.ru**

**Хлопотов Максим Валерьевич, канд. техн. наук, доцент, факультет инфокоммуникационных технологий, ИТМО, khlopotov@itmo.ru**

Computer tools in education, 2024
№ 1: 32–44
http://cte.eltech.ru
doi:10.32603/2071-2340-2024-1-32-44

# Exploring the Effectiveness of Prompt Engineering and Quantized Large Language Models in the Development of Academic Courses

Shnaider P. A. [1], Postgraduate, Assistant, beatrix.linkoln@gmail.com, orcid.org/0000-0002-4147-6561
Chernysheva A. V.[1], Assistant, ✉ avchernysheva@itmo.ru, orcid.org/0000-0002-9956-6607
Nikiforova A. D.[1], Student, 34743@niuitmo.ru
Govorov A. I.[1], Senior Lecturer, govorov@itmo.ru, orcid.org/0009-0005-6674-1666
Khlopotov M. V.[1], Cand. Sc., Associate Professor, khlopotov@itmo.ru, orcid.org/0000-0002-9053-027X

[1]ITMO University, 49 Kronverksky, bldg. A, 197101, Saint Petersburg, Russia

**Abstract**

Abstract. This article presents the outcomes of an experiment employing large language models (LLMs) in the development of university course structures. Various prompt

engineering methods, including zero-shot, few-shot, chain-of-thought, and tree-of-thought, were employed to formulate queries to LLMs. Primarily, quantized models such as mistral-7b-instruct, mixtral-8x7b-instruct, openchat_3.5, saiga2_13b, starling-lm-7b-alpha, tinyllama, among others, were utilized for the experiment. The generated course structures were compared with data obtained from ChatGPT-4. Models openchat_3.5.q5_k_m and starling-lm-7b-alpha.q5_k_m demonstrated comparable quality in generating educational program structures to ChatGPT-4. The experiment underscores the potential applications of LLMs in the field of education and highlights promising directions for further research.

**Keywords:** *Large Language Models, Prompt Engineering, Quantized Models, few-shot, zero-shot, chain-of-thought*

**Citation:** P. A Shnaider., A. V. Chernysheva, A. D. Nikiforova, A. I. Govorov, and M. V. Khlopotov, "Exploring the Effectiveness of Prompt Engineering and Quantized Large Language Models in the Development of Academic Courses," *Computer tools in education*, no. 1, pp. 32–44, 2024 (in Russian); doi:10.32603/2071-2340-2024-1-32-44

## References

1. E. Jermakowicz, "The Coming Transformative Impact of Large Language Models and Artificial Intelligence on Global Business and Education," *Journal of Global Awareness*, vol. 4, no. 2, pp. 1–22, 2023; doi:10.24073/jga/4/02/03
2. S. Laato, B. Morschheuser, J. Hamari, and J. Björne, "AI-Assisted Learning with ChatGPT and Large Language Models: Implications for Higher Education," in *2023 IEEE International Conference on Advanced Learning Technologies (ICALT), Orem, Utah, United States*, pp. 226–230, 2023; doi:10.1109/icalt58122.2023.00072
3. A. Giretti et al., "Integrating large language models in art and design education," in *Proc. of the Int. Conf. on Cognition and Exploratory Learning in the Digital Age, Madeira Island, Portugal, 21-23 October, 2023*, pp. 1–7, 2023.
4. J. Jeon and S. Lee, "Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT," *Education and Information Technologies*, vol. 28, no. 12, pp. 15873–15892, 2023; doi:10.1007/s10639-023-11834-1
5. M. Abedi, I. Alshybani, M. Shahadat, and M. Murillo, "Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education," in *qeios.com*, 2023 [Preprint]; doi:10.32388/md04b0.2
6. J. Prather et al., "The Robots Are Here: Navigating the Generative AI Revolution in Computing Education," in *Proc. of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education*, 2023; doi:10.1145/3623762.3633499
7. M. Irfan and L. Murray, *Micro-Credential: A Guide to Prompt writing and Engineering in Higher Education: A tool for Artificial Intelligence in LLM*, Limerick, Ireland: University of Limerick, 2023; doi:10.13140/RG.2.2.15596.95367
8. T. Kojima, S. (Shane) Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large Language Models Are Zero-Shot Reasoners," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 22199–22213, 2022.
9. P. Dhariwal et al., "Language Models Are Few-Shot Learners," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*, vol. 33. pp. 1–25, 2020.
10. J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 1–14, 2022.
11. S. Yao et al., "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," in *arxiv.org*, 2023. [Online]. Available: https://arxiv.org/abs/2305.10601
12. S. Li et al., "LLM-MQ: Mixed-precision Quantization for Efficient LLM Deployment," in *NeurIPS 2023 Efficient Natural Language and Speech Processing Workshop*, pp. 1–5, 2023.
13. W. Li, A. Hu, N. Xu, and G. He, "Quantization and Hardware Architecture Co-Design for Matrix-Vector Multiplications of Large Language Models," *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 1–14, 2024; doi:10.1109/tcsi.2024.3350661

14. A. Q. Jiang et al., “Mixtral of Experts,” in *arxiv.org*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.04088
15. G. Wang et al., “OpenChat: Advancing Open-source Language Models with Mixed-Quality Data,” in *arxiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.11235
16. B. Zhu et. al, “Starling-7B: Improving LLM Helpfulness & Harmlessness with RLAIF,” in *starling.cs.berkeley.edu*, 2023. [Online]. Available: https://starling.cs.berkeley.edu/
17. C. Irugalbandara, “A Trade-off Analysis of Replacing Proprietary LLMs with Open Source SLMs in Production,” in *arxiv.org*, 2024. [Online]. Available: arXiv:2312.14972https://doi.org/10.48550/arXiv.2312.14972
18. M. Tikhomirov and D. Chernyshev, “Impact of Tokenization on LLaMa Russian Adaptation,” in *arxiv.org*, 2023. [Online]. Available: https://arxiv.org/html/2312.02598v1
19. J. Chen, H. Lin, X. Han, and L. Sun, “Benchmarking Large Language Models in Retrieval-Augmented Generation,” in *arxiv.org*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.01431

*Received 29-01-2024, the final version — 29-02-2024.*

**Polina Shnaider, Postgraduate, Assistant, Faculty of Infocommunication Technologies, ITMO University, beatrix.linkoln@gmail.com**

**Anastasiia Chernysheva, Assistant, Faculty of Infocommunication Technologies, ITMO University, ✉ avchernysheva@itmo.ru**

**Anna Nikiforova, 3rd year Student of the bachelor’s degree program, Faculty of Infocommunication Technologies, ITMO University, 34743@niuitmo.ru**

**Anton Govorov, Senior Lecturer, Faculty of Infocommunication Technologies, ITMO University, govorov@itmo.ru**

**Maksim Khlopotov, Candidate of Sciences (Tech.), Associate Professor, Faculty of Infocommunication Technologies, ITMO University, khlopotov@itmo.ru**