INTERFACE
–JOURNAL OF EUROPEAN LANGUAGES AND LITERATURES
ISSN: 2519-1268

Issue 26 (Spring 2025), pp. 195-232
DOI: 10.6667/interface.26.2025.257

Опыт дообучения языковой модели – от бытового языка к философскому

ДМИТРИЙ А. ЯРОЧКИН
Санкт-Петербургский государственный университет

Аннотация

Статья исследует интеграцию философии, в частности аристотелевской мысли, в сферу искусственного интеллекта В ней описывается проект, направленный на дообучение языковой модели с целью повышения её способности обрабатывать и генерировать философские тексты. Основная цель заключается в улучшении понимания моделью сложных философских концепций и её способности к критическому мышлению. Это достигается за счёт тонкой настройки модели на корпусе текстов Аристотеля и связанных с ними аналитических материалов. Исследование использует как количественные метрики, так и качественную экспертизу для оценки производительности модели, включая её способность работать с этическими и онтологическими вопросами. В конечном итоге, работа направлена на развитие более объяснимых и этически обоснованных систем искусственного интеллекта, использующих философские подходы для улучшения способности ИИ к рассуждению и концептуальному пониманию

Ключевые слова: Интеграция философии, Аристотель, языковая модель, искусственный интеллект, философские тексты, этика и онтология

INTERFACE
–JOURNAL OF EUROPEAN LANGUAGES AND LITERATURES
ISSN: 2519-1268

Issue 26 (Spring 2025), pp. 195-232
DOI: 10.6667/interface.26.2025.257

Experience of Fine-Tuning a Language Model – From Everyday Language to Philosophical Language

DMITRIY A. YAROCHKIN
Saint Petersburg State University

Abstract

The article explores the integration of philosophy, specifically Aristotle's thought, with artificial intelligence. It details a project focused on fine-tuning a language model to enhance its ability to process and generate philosophical texts. The major aim is to improve the model's understanding of complex philosophical concepts and its capacity for critical thinking. This is achieved through fine-tuning the model using a corpus of Aristotle's works and related research. The study employs both quantitative metrics and qualitative expert evaluations to assess the model's performance, including its ability to handle ethical and ontological questions. Ultimately, the research aims to contribute to the development of more explainable and ethically grounded AI systems, using philosophical frameworks to improve AI's reasoning and conceptual comprehension.

Keywords: Philosophy Integration, Aristotle, Language Model, Artificial Intelligence, Philosophical Texts, Ethics and Ontology.

Опыт дообучения языковой модели – от бытового языка к философскому

1 Введение. Интеграция философии в модель ИИ: проблемы и возможности для объяснимого ИИ

Проблема, с которой столкнулся проект «лаборатория цифровой философии» заключается в том, что модели глубокого обучения, такие как RuGPT-3, эффективны для обработки текстов, но сталкиваются с трудностями при интерпретации сложных философских концепций и терминов. Философские тексты содержат специфические термины и логические структуры, которые модели часто воспринимают поверхностно, опираясь на ключевые слова, а не их контекстуальное значение, что может приводить к ошибкам в интерпретации.

Поверхностное освоение философской терминологии Искусственным Интеллектом (далее «ИИ») может заменить философский язык упрощенными конструкциями, что представляет риск для точности философского анализа. Философия отражает фундаментальные основания человеческого существования, выраженные в таких науках как этика, антропология и онтология. Глубокое освоение этих категорий ИИ важно для создания объяснимого ИИ (XAI) (Gunning & Aha, 2019), в котором мотивация решений будет прозрачной, и для повышения безопасности и этичности ИИ-систем.

Дообучение модели ИИ на текстах Аристотеля может повысить **прозрачность** (понимание пользователями принципов работы ИИ) и **объяснимость** (способность ИИ разъяснять свои решения) за счет структурированной логики, категориального аппарата, риторики и этики. Аристотелевская логика и строгая система

категорий помогают ИИ формулировать более точные и обоснованные выводы, избегая упрощенных конструкций. Этика и антропология способствуют учету человеческих ценностей при объяснении решений, а риторика развивает способность ясно и убедительно излагать информацию. Таким образом, интеграция философских принципов делает XAI более понятным, логичным и достойными доверия пользователей.

Для реализации этой задачи нам надо ответить на следующие вопросы:

1. Как дообучение модели RuGPT-3 на корпусе философских текстов может улучшить интерпретацию сложных философских концепций и терминов, снижая риск поверхностного восприятия?
2. Может ли дообучение модели на философских текстах, таких как труды Аристотеля, способствовать более глубокому освоению философской терминологии и улучшению контекстуального понимания текстов?
3. Каким образом дообучение модели на философских текстах может способствовать развитию этичного и объяснимого ИИ, где мотивация решений будет прозрачной и понятной?

Эти вопросы направлены на решение проблемы поверхностного восприятия философских текстов моделями глубокого обучения и исследуют, как дообучение может способствовать более глубокому и точному пониманию философских концепций. Отвечая на них, мы планируем проверить следующие гипотезы.

1.1 Рабочие гипотезы исследования:

1. Дообучение модели RuGPT-3 на корпусе философских текстов повысит ее способность генерировать релевантные, концептуально точные тексты, соответствующие

философской традиции.
2. Дообучение модели на философских текстах способствует формированию у нее более глубокой обработки семантических связей, позволяя ей воспринимать текст в его целостности, а не через отдельные ключевые слова. Это может повышать уровень абстракции модели и ее способность к интерпретации философских концептов.
3. Данный процесс представляет собой этап на пути к разработке объяснимых систем искусственного интеллекта, способных к осмысленному воспроизведению и интерпретации философских категорий и дискурсов.

1.2 Цели исследования:

1. Повышение релевантности и точности текстов, сгенерированных RuGPT-3, в контексте философских вопросов и концептов, включая классическую философию и современные философские дискуссии.
2. Разработка методик дообучения (fine-tuning) для моделей в гуманитарных дисциплинах, с фокусом на адаптацию моделей под специфические требования и сложности философских текстов.
3. Адаптация и интеграция объяснительной модели Аристотеля для повышения способности модели к философским рассуждениям, что способствует более глубокому пониманию и воспроизведению философских категорий и методов анализа.

Предложенный набор целей и вопросов направлен на всестороннее исследование возможности адаптации RuGPT-3 к философским задачам и применения модели в академических и образовательных контекстах, а также на оценку её потенциала в более широком контексте гуманитарных наук. Поскольку современные технологии дообучения предлагают множество подходов, необходимо тщательно выбрать метод, который наиболее эффективно

соответствует задаче освоения моделью философского стиля мышления.

2 Различные модели обучения ИИ. Проблема прозрачности моделей

В данном исследовании для обработки длинных текстов Аристотеля выбор модели RuGPT-3 обоснован её способностью эффективно захватывать долгосрочные зависимости и поддерживать контекст фразы, что критически важно для философских рассуждений. В отличие от рекуррентных нейронных сетей (RNN) и их улучшенных версий, таких как LSTM, которые могут терять смысл на длинных участках текста, трансформеры обеспечивают более точную генерацию текста и учитывают глобальный контекст. Таким образом, трансформеры являются оптимальным выбором для задач, требующих сохранения смысловых связей и контекста в длинных текстах.

Однако с ростом мощности моделей ИИ, падает их прозрачность, то есть то, насколько пользователь может понять принципы работы модели и обоснование ее решений. Следующий график иллюстрирует данное явление.

Многослойность архитектуры моделей приводит к тому, что мы не можем понять ее мотивацию. Для преодоления проблемы черного ящика, которую порождает система искусственного интеллекта, разработаны различные ad-hoc методики анализа, в том числе и текстуального. «[Её] внутреннее устройство остаётся загадкой для пользователей. Пользователи могут видеть входные данные и результаты работы системы, но не могут понять, что происходит внутри инструмента ИИ, чтобы эти результаты были получены» (Arrieta at al, 2020).

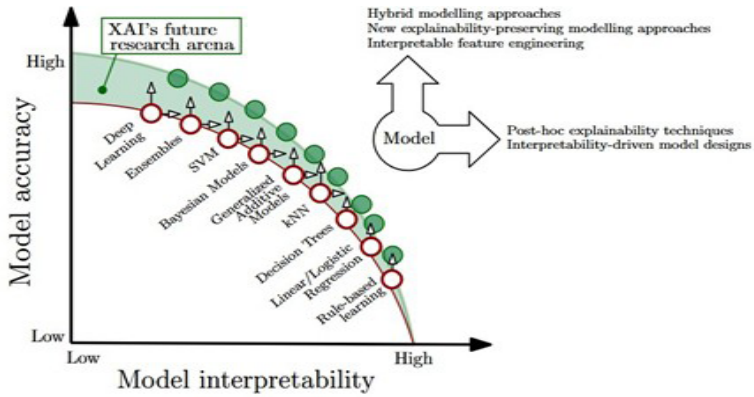

Figure 12: Trade-off between model interpretability and performance, and a representation of the area of improvement where the potential of XAI techniques and tools resides.

Рисунок 1. Соотношение объяснимости и точности моделей (Arrieta at al, 2020)

Применение философских концепций и методов к ИИ может повысить способность ИИ решать сложные проблемы, принимать этически обоснованные решения и способствовать более полной интеграции ИИ в различные области.

Нам кажется, что философия Аристотеля может внести вклад не только в теоретические области знания с нею связанные, но и в практические. ИИ может внести вклад в процесс образования, становясь тренажером для написания текстов, улучшения риторики, повышения интерактивности занятий и преодоления страха перед выступлениями. Это, в свою очередь, может оказать положительное влияние на сферу философского образования и философского знания в целом, вводя новые методики. Более практичными выводами можно считать описанную выше попытку углубить понятийный аппарат объяснимого ИИ, т.е., внедрить встроенные принципы этики и построить более безопасный и человекоориентированный ИИ. В конечном итоге это сказывается как на доверии общества к технологии, так и на её безопасности в долгосрочной перспективе.

3 Философия Аристотеля как основа прозрачности ИИ.

Доктрина Аристотеля о четырех причинах утверждает, что Благо является причиной бытия (Aristotle, 2020). Его идея о разделении формы и материи позволяет переосмыслить природу искусственного интеллекта (Корниенко и др., 2013), тогда как платоновский дуализм, предполагающий нематериальность разума (Aristotle, 2020), менее применим в техническом контексте. Благо – это то, к чему стремятся все вещи (Aristotle, 2020), и, в связи с этим возникает вопрос, должен ли искусственный интеллект также стремиться к Благу. Размышляя о душе, Аристотель выделяет несколько ее частей, из которых только рациональная часть может быть отделена (Aristotle, 1907). Отсутствие «души» у искусственного интеллекта не исключает возможность этики.

Если понимать этику как стремление к благу, то искусственный интеллект, как и любой другой объект в мире, подчиняется этому стремлению. Этот подход не только морален, но и выгоден в долгосрочной перспективе, поскольку понятие блага выходит за рамки экономической выгоды или мимолетного удовольствия (Sparks & Wright, 2023) и связано с долгосрочными выгодами (Brennan-Marquez & Henderson, 2019).

Внедрение философских категорий в процесс дообучения способствует созданию объяснимых выводов и более прозрачных решений, что позволяет исследователям и практикам легче интерпретировать и объяснять поведение модели. В данном контексте, философские тексты, насыщенные сложными абстракциями и логическими структурами, могут стать важным инструментом для разработки более прозрачных и понятных механизмов работы ИИ.

3.1 Методы XAI

В последние годы наблюдается экспоненциальный рост использования глубокого обучения (DL) в различных областях, включая, например, здравоохранение. Однако, несмотря на их впечатляющую производительность, модели глубокого обучения, такие как глубокие нейронные сети (DNN), часто рассматриваются как «черные ящики» из-за своей сложности и непрозрачности (Salvi et al., 2025). Это создает серьезные проблемы. Люди неохотно принимают технологии, которые они не могут понять и которым не могут доверять (Chazette & Schneider, 2020; Salvi et al., 2019). В связи с этим, область объяснимого ИИ (XAI) приобретает все большее значение, предлагая методы для создания более прозрачных и понятных моделей ИИ, сохраняя при этом высокую точность прогнозирования.

Существует несколько подходов к достижению объяснимости в ИИ. В статье Л. Чжан, посвященной тому, как XAI может быть применен для анализа данных, связанных с геотермальными исследованиями (Zhang et al., 2025), рассматриваются две основные стратегии: упрощение и определение значимости признаков. В данной работе для объяснения модели используется метод Шеплиевых аддитивных объяснений (SHAP)[1], который позволяет получить представление о взаимосвязи между входными и выходными данными модели, повышая прозрачность и надежность прогнозов. Другие работы также подчеркивают важность SHAP и LIME[2] для анализа значимости признаков (Iqbal et al.,

1 **Шеплиевые аддитивные объяснения**, или **SHapley Additive exPlanations (SHAP) –** метод, который основывается на идее, что вклад каждого признака (или переменной) можно оценить с учетом его влияния на модель в разных возможных комбинациях с другими признаками. Суть метода заключается в вычислении Шеплиевых цен для каждого признака, которые отражают его средний вклад в решение модели по сравнению с другими признаками.

2 **Локальные интерпретируемые объяснения, независимые от модели**, или LIME (Local Interpretable Model-agnostic Explanations) – механизм построения локальной интерпретируемой модели на основе небольшого набора данных, который аппроксимирует поведение сложной модели только для конкретного случая. Это делает объяснение доступным и понятным для пользователя, несмотря на сложность самой модели.

2025; Benlecheb et al., 2025).

4 Использование ИИ в цифровой гуманитаристике

Методы объяснимого искусственного интеллекта (XAI), такие как SHAP, играют ключевую роль в обеспечении прозрачности и интерпретируемости моделей ИИ, что особенно важно в гуманитарных исследованиях. В цифровой гуманитаристике, где анализ текстов, интерпретация смыслов и моделирование сложных концепций требуют глубокого понимания, XAI помогает преодолеть ограничения «чёрного ящика». Это позволяет исследователям не только доверять результатам, но и использовать ИИ как инструмент для изучения философских, исторических и культурных феноменов, обеспечивая более точное и осмысленное взаимодействие между технологиями и гуманитарными дисциплинами.

Интеграция ИИ в цифровую гуманитаристику открывает новые возможности для исследований и обучения. Теоретико-ориентированная наука о данных (Karpatne et al., 2017) объединяет философские принципы с анализом данных, что способствует созданию более содержательных и интерпретируемых моделей ИИ. Философия также играет ключевую роль в проектировании систем, ориентированных на человека, обеспечивая учёт этических норм и общественного благополучия (Oxford AI Ethics, 2024).

Искусственный интеллект (ИИ) может играть роль философской лаборатории для мысленных экспериментов, однако современные языковые модели часто подвергаются критике за их поверхностное понимание смысла (Rees, 2022). Модель INTUITEL, основанная на семантическом вебе, является примером философского бота (Verdú et al., 2017). Тем не менее, внедрение ИИ в педагогическую практику в области философии остается недостаточно исследованным аспектом.

Продолжаются дискуссии о преимуществах и недостатках внедрения ИИ в образовательный процесс, поскольку это связано с рядом экзистенциальных рисков, включая возможность утраты человеком критических навыков. В контексте философии данная угроза обостряется, поскольку философия представляет собой систему абстрактных принципов мышления, формирующую горизонт понимания для всего человечества. Утрата автономности человека в данной сфере может привести к упрощению мысли, но в то же время ИИ предлагает философии новые возможности, становясь лабораторией для мысленного эксперимента.

В образовании ИИ способствует переходу от подхода, ориентированного на результат, к процессуальному, поддерживая такие аспекты как персонализацию обучения, автоматическую оценку и улучшение навыков письма (Velbor Bazic, Indrasen Poola; Lin & Chang, 2020). Технологии также позволяют моделировать процесс обучение со стороны ученика, что позволяет добиться определённой доли адаптивности образовательного процесса (Lee et al., 2024), хотя есть и ограничения, касающиеся того, насколько модели релевантны тому, что они моделируют, – действительно мышление ученика. Одним из перспективных направлений исследований можно назвать метод моделирования ошибок студентов (generative student error modeling) – это подход в интеллектуальных обучающих системах, который предсказывает ошибки учащихся, анализируя их когнитивные процессы и искажения знаний. Он используется для адаптивного обучения, однако подвергается критике за фокус на ошибках вместо глубинных проблем обучения, таких как семантические искажения и поверхностная обработка информации. Глобальное исследование проблем использования ИИ в образовании важно для улучшения образовательных практик (Leach, 2008). Кроме того, ИИ может выступать в роли культурного посредника, улучшая атмосферу в классе (Kim, 2016), и поддерживать динамическую адаптацию к индивидуальным потребностям студентов (Wasson, 1998).

Процесс создания и обучения нейросетей предоставляет учащимся

уникальную возможность для практического осмысления их устройства, что способствует лучшему пониманию работы моделей и формированию системы прозрачности для обеспечения доверия общества к новым технологиям (Ларионов & Ярочкин, 2024). ИИ становится важным инструментом в гуманитарных дисциплинах, способствуя как пониманию, так и созданию нового знания, хотя его ограничения требуют дальнейшего изучения и интеграции с философскими и этическими принципами. Хотелось бы отметить, что эвристическая функция ИИ остаётся спорной: неясно, создает ли ИИ новое знание или комбинирует старое, то есть, являются ли знанием комбинация и установление скрытых связей. Скорее, мы можем утверждать, что применение технологий ИИ позволит исследователю приходить к неожиданным выводам, чем повышает его продуктивность; и в этом смысле ИИ выступает помощником человека. Как указано в сборнике тезисов «The Lyceum Project: Ai Ethics With Aristotle», говорящим орудием (Ober, J., & Tasioulas, J.. p 3), а не руководителем. Именно в установлении такой иерархии лежит большая часть стратегий преодоления рисков ИИ в образовании. Как отметили руководители проекта «Лаборатория цифровой философии», в этом смысле использование ИИ ничем не отличается от использования редакторов текста при его наборе.[3] Главное ограничение, которое препятствует творческому использованию ИИ, можно увидеть в самой технологии языковых моделей. Они повторяют распределение слов и выдают сглаженную, верную, среднестатистическую речь. С этой точки зрения они противоположны речи философской. Яркими примерами философский речи могут быть речь Гераклита, прозванного «темным» за неясность слога, или Хайдеггера. Тексты и того, и другого наглядно демонстрируют, что человеческий язык не приспособлен для точного выражения философской идеи. Поэтому философы вынуждены создавать свой собственный язык, ломающий привычные схемы. Так или иначе такая «ломка», уточнение языка, является типичной для философии. Философия часто берет привычные бытовые термины и наделяет их

3 Актуальный репортаж: ChatGPT напишет курсовую? https://rutube.ru/video/0aab383d201b511a9e5dff428c01fca8/

глубокими смыслами и взаимосвязями, которые в обычном языке опускаются. Это философское уточнение становится критичным, например, в этике, поскольку формулирует отношение к обществу через поступок, который является единичным явлением, а не статистическим фактом.

Для начала преодоления указанных ограничений предлагается дообучение модели на текстах Аристотеля, что может повысить её способность к генерации релевантных и концептуально точных текстов. Философия Аристотеля, с её акцентом на логические структуры и абстрактные категории, такие как благо, справедливость и этика, предоставляет уникальную основу для развития более прозрачных и объяснимых моделей ИИ. Это не только улучшает качество генерации текстов, но и способствует созданию систем, способных к осмысленному воспроизведению философских дискурсов, что открывает новые возможности для применения ИИ в образовании, академических исследованиях и других гуманитарных дисциплинах.

4.1 Методы

В данной работе авторы исследуют процесс интеграции философских концепций, в частности, аристотелевской мысли, в разработку моделей искусственного интеллекта. Методы исследования включают несколько ключевых этапов обработки данных и оценки философской релевантности моделей. На начальном этапе проводится отбор текстов экспертами, что позволяет выявить наиболее значимые и релевантные материалы для анализа. После этого осуществляется очистка и нормализация данных, направленная на устранение шума и приведение данных к стандартизированному виду, после чего формируется датасет в формате CSV, обеспечивающий удобство дальнейшей работы. Для выделения скрытых тем и формирования релевантных промптов используется метод Latent Dirichlet Allocation (LDA) (Blei at al., 2003), который помогает определить ключевые темы в текстах.

4.2 Архитектура базовой модели

В рамках нашего проекта мы использовали модель «ruGPT3small_based_on_GPT2», (Zmitrovich et al., 2024). Эта модель (125 млн параметров) обучалась на 450 ГБ текстов (Википедия, новости, книги, Colossal Clean Crawled Corpus, OpenSubtitles) с длиной последовательности 1024 токена (80 миллиардов токенов, 3 эпохи), затем дообучалась с длиной 2048 токенов за 1 эпоху; общее время обучения составило 14 дней на 128 GPU для контекста 1024 и несколько дней на 16 GPU для контекста 2048.

4.3 Методы дообучения и оценки

Для улучшения качества модели применяется дообучение без учителя, что позволяет повысить точность генерации текстов без необходимости разметки данных. В исследовании выполнено полное дообучение модели ruGPT-3 Small адаптации к философскому стилю на корпусе текстов Аристотеля[4], что является

4 Код дообучения на языке python3:
from datasets import load_dataset from transformers import (AutoTokenizer, AutoModelForCausalLM, TrainingArguments, Trainer, DataCollatorForLanguageModeling, EarlyStoppingCallback)
MODEL_NAME = “ai-forever/rugpt3small_based_on_gpt2” OUTPUT_DIR = “aristotle_csv2”
MODEL_HUB_NAME = “aristotle_csv2”
dataset = load_dataset(‘DmitryYarov/aristotle_csv’)
train_test_dataset = dataset[‘train’]. train_test_split(test_size=0.1, seed=42) train_dataset = train_test_dataset[‘train’]
eval_dataset = train_test_dataset[‘test’]
tokenizer = AutoTokenizer.from_pretrained (MODEL_NAME, use_fast=True) tokenizer.pad_token = tokenizer.eos_token
model = AutoModelForCausalLM.from_pretrained (MODEL_NAME) def tokenize_function(examples):
return tokenizer (examples[“text”], truncation=True, max_length=512, padding=”max_length”
tokenized_train_dataset = train_dataset.map(tokenize_function, batched=True, remove_columns=[“text”]) tokenized_eval_dataset = eval_dataset.map(tokenize_function, batched=True, remove_columns=[“text”]) data_collator = DataCollatorForLanguageModeling(tokenizer=tokenizer, mlm=False)
training_args = TrainingArguments(output_dir=OUTPUT_DIR, evaluation_strategy=”epoch”, save_strategy=”epoch”, save_total_limit=3, # Save only the 3 most recent checkpoints logging_dir=f”{OUTPUT_DIR}/logs”, logging_steps=50, per_device_train_batch_size=8, per_device_eval_batch_size=8, num_train_epochs=30, weight_decay=0.01, learning_rate=5e-5, warmup_steps=500, gradient_accumulation_steps=4, gradient_checkpointing=True, fp16=True, optim=”adafactor”, push_to_hub=True, report_to=”none”, # Disable unwanted logs load_best_model_at_end = True,)
trainer = Trainer (model=model, args=training_args, train_dataset=tokenized_train_dataset, eval_dataset=tokenized_eval_dataset, data_collator=data_collator callbacks=[EarlyStoppingCallback(early_

примером «Domain Adaptation». Использован специализированный «Датасет» (пояснение о нем см.: раздел «Данные»), настроен токенизатор с eos_token как pad_token. Тексты токенизированы с максимальной длиной 512 токенов и добавлением padding. Обучение проводилось с параметрами: 30 эпох, размер батча 8, Gradient Accumulation 4, FP16 для ускорения, оптимизатор Adafactor, EarlyStopping через 3 эпохи без улучшений, автосохранение – 3 последних чекпоинта.

Для оценки хода дообучения мы использовали функцию потерь[5], чтобы измерить отклонение предсказаний модели от эталонных данных, что позволило отслеживать улучшение точности и терминологической адаптации. Дополнительно применялись метрики, такие как BERTScore[6], для оценки семантической релевантности и логической целостности генераций. Также для анализа важности признаков в генерациях использовался метод SHAP (Shapley Additive Explanations), который выявляет влияние отдельных характеристик модели на её предсказания, помогая глубже понять, какие аспекты данных влияют на результаты. Наконец, метод Expert-in-the-Loop позволил провести качественную экспертную оценку генераций, где эксперт в соответствующей предметной области анализировал тексты, что обеспечивало более детализированную оценку соответствия философским стандартам и возможность выявления проблем с логической целостностью (Securly AI Chat, 2025).

---

stopping_patience=3)],)

trainer. train ()

5 **Loss** (потери) – это функция, измеряющая отклонение предсказаий модели от реальных значений. Чем меньше потери, тем точнее модель. В процессе дообучения модель минимизирует функцию потерь, корректируя свои параметры для улучшения точности предсказаний и адаптации к специфике данных.

6 **BERTScore** – это метрика для оценки качества текстов, основанная на моделях BERT, которая измеряет сходство между предсказанными и эталонными текстами (у нас это тексты Аристотеля, вошедшие в «Датасет») на основе их представлений в многослойных трансформерах. BERTScore использует контекстуальные векторные представления слов для вычисления сходства между текстами, что позволяет более точно оценить качество перевода или генерации текста, учитывая семантические и синтаксические аспекты.

4.4 Данные

В корпус текстов, на которых обучалась модель, вошли следующие издания Аристотеля на русском языке:

1. Аристотель (1976). Собрание сочинений в 4-х томах. Т. 1. Москва: Мысль.
2. Аристотель (1978). Метафизика, о душе. Собрание сочинений в 4-х томах. Т. 2. Москва: Мысль.
3. Аристотель (1981). Категории, о истолковании, первая аналитика, вторая аналитика, книга первая, книга вторая, топика, о софистических опровержениях. Собрание сочинений в 4-х томах. Т. 3. Москва: Мысль.
4. Аристотель (1983). Физика, о небе, о возникновении и уничтожении, метеорологика. Собрание сочинений в 4-х томах. Т. 4. - Аристотель. Никомахова этика, Политика.

Кроме того, были включены следующие философские работы: «Метафизика Аристотеля» (В.Ф. Асмус); «Основоположения логики Аристотеля» (Г.Д. Микеладзе); «Естественнонаучные сочинения Аристотеля» (Рожанский И.Д.); «Этические сочинения Аристотеля» (Кессиди Т.Ф.); «Политика Аристотеля» (Доватур А.И.).

Целью такого подхода было обеспечить контекстуальную релевантность русского языка. Для подготовки данных был проведён процесс их очистки и нормализации, после чего они были собраны в датасет в формате CSV (в тексте статьи: «Датасет»). Каждое предложение из текстов было представлено как отдельная сущность в Датасете, общее количество которых составило 7187.

4.5 Результаты

4.5.1 График сходимости функции потерь при дообучении

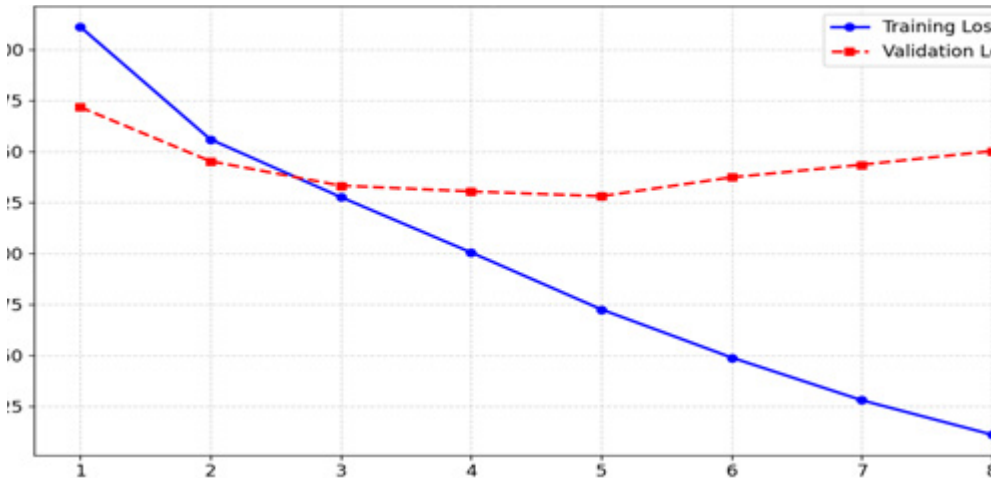

Рисунок 2. График сходимости функции потерь при дообучении

На графике сходимости функции потерь показана динамика потери на обучение (Training Loss) и потери на валидацию (Validation Loss) в процессе дообучения модели. Показатель потерь на обучение стабильно снижается, что свидетельствует об успешном изначальном обучении, однако рост потерь на валидацию после 3-5 эпохи указывает на процесс переобучения, когда модель начинает запоминать детали обучающих данных, теряя способность эффективно обобщать информацию при использовании новых данных. Для предотвращения этого был применён метод ранней остановки[7], сохранивший оптимальное состояние модели.

7 **Метод ранней остановки** (early stopping) – техника, используемая для предотвращения переобучения, при которой обучение модели прекращается, если показатель **Validation Loss** перестает улучшаться на протяжении определённого числа эпох. Это позволяет сохранить оптимальное состояние

4.5.2 Оценка генерации текста: интеграция BERTScore, SHAP и экспертная оценка («Expert-in-the-Loop») для объяснимого ИИ (XAI)

Анализ графика потерь показал, что модель успешно обучается, но требует контроля для предотвращения переобучения. Однако для полноценной оценки её способностей, особенно в контексте генерации философских текстов, необходимо не только проанализировать процесс обучения, но и разработать инструменты для проверки качества и релевантности сгенерированных результатов. В связи с этим был сформирован стандарт оценки модели, основанный на тематическом анализе текстов Аристотеля, а также были добавлены ключевые для статьи философские концепции, связанные с творчеством философа (сознание, философский метод, концепция этики говорящего орудия), которые мы добавили для сравнения и анализа гибкости системы. Этот стандарт позволяет систематически оценивать, насколько модель способна генерировать тексты, соответствующие аристотелевской философии.

4.5.2.1 Формирование стандарта для оценки модели

Для оценки модели на основе философских концепций Аристотеля был проведён тематический анализ с применением метода латентного размещения Дирихле (LDA)[8]. В результате были выделены 9 ключевых тем, характеризующих тексты Аристотеля. На основе этих принципов был сформирован бенчмарк, включающий набор промптов, предназначенных для оценки способности модели генерировать тексты, соответствующие философским

модели, избегая её переобучения на тренировочных данных. (см. https://huggingface.co/docs/transformers/en/main_classes/callback#transformers.EarlyStoppingCallback)

8 **LDA (Latent Dirichlet Allocation)** – статистическая модель, используемая для тематического моделирования текстов, которая предполагает, что каждый документ является смесью различных скрытых тем, каждая из которых представляет собой распределение слов. Модель обучается на данных, извлекая скрытые темы с использованием таких методов как Гиббсовая выборка.

концепциям, содержащимся в корпусе аристотелевских произведений. В ходе тематического моделирования методом LDA были выделены 9 тем, характеризующих тексты Аристотеля.

На основании полученных результатов по частотам терминов были выведены следующие промпты:

Рисунок 3. Пузырьковая диаграмма по частотам терминов в темах.

Таб. 1. Промпты исследования

Промпты	**Определение термина в первоисточнике**	**Значение термина для исследования**
Благо – это	«Всякое искусство и всякое учение, а равный образом поступок (praxis) и сознательный выбор, как принято считать, стремятся к определенному благу. Поэтому удачно определяли благо как то, к чему все стремится». (Аристотель, 1983, с. 54)	В отношении задач статьи такой подход позволяет говорить о Благе для неодушевленных предметов, таких как ИИ.

Добродетель – это	«Итак, добродетель есть сознательно избираемый склад [души], состоящий в обладании серединой по отношению к нам, причём определённой таким суждением, каким определит её рассудительный человек. Серединой обладают между двумя [видами] порочности, один из которых – от избытка, другой – от недостатка». (Аристотель, 1983, с. 87)	Понятие добродетели, ппредложенное Аристотелем, содержит потенциал для урегулирования безопасности ИИ
Сущее – это	«Итак, сущее и единое – одно и то же, и природа у них одна, поскольку они сопутствуют друг другу так, как начало и причина». (Аристотель, 1976, с. 119)	Важно для исследования, поскольку отражает фундаментальную категорию метафизики, связанную с бытием и сущностью.
Бытие – это	«Бытие же само по себе приписывается всему тому, что обозначается через формы категориального высказывания, ибо сколькими способами делаются эти высказывания, в стольких же смыслах обозначается бытие». (Аристотель, 1976, с. 156)	Аристотель предложил десять категорий бытия: сущность, качество, количество, отношение, место, время, положение, обладание, действие, страдание. Эти категории выражают различные смыслы бытия, зависящие от формы высказывания, что позволяет анализировать онтологические характеристики объектов, включая современные технологические артефакты.

Материя – это	«Естественным путем, стало быть, существует то, что состоит из материи и формы, например живые существа и части их тела; а естество – это, с одной стороны, первая материя, с другой стороны, форма, или сущность». (Аристотель, 1976 с. 150)	Материя противопоставляется форме и является потенциалом для становления сущности. Все вместе эти три промпта отражают различные аспекты в метафизике Аристотеля. Такая аналитика не свойственна обыденному языку, поэтому эти промпты могут служить хорошей проверкой понимания сути текстов Аристотеля.
Возможности – это	«Способностью, или возможностью (dynamis), называется начало движения или изменения вещи, находящееся в ином или в ней самой, поскольку она иное». (Аристотель, 1976, с. 162)	Противопоставляется актуальности и рассматривается как потенциал изменений, что тоже отлично от обыденного языка.

Душа – это	«душа есть первая энтелехия естественного тела, обладающего органами. Потому и не следует спрашивать, есть ли душа и тело нечто единое, как не следует это спрашивать ни относительно воска и отпечатка на нем, пи вообще относительно любой материи и того, материя чего она есть. Ведь хотя единое и бытие имеют разные значе ния, но энтелехпя есть единое и бытие в собственном смысле». (Аристотель, 1976, с. 395)	Отражает связь души и тела, важную для управления безопасностью ИИ.
Знание – это	«Совершенно очевидно, что необходимо приобрести знание о первых причинах: ведь мы говорим, что тогда знаем в каждом отдельном случае, когда полагаем, что нам известна первая причина». (Аристотель, 1976, с. 25)	Отличное от поверхностного понимания знание, как знание метафизики, показывает, насколько модель стремится глубоко анализировать промпты. Оно показывает глубину понимания и объяснения.
Справедливость – это	«Справедливое по отношению к другому есть, собственно говоря, равенство». (Аристотель, 1983, с. 324)	Фундаментальная аналитика справедливости может приобретать количественную меру и в этом смысле являться основой аналитики вводных данных модели.

Не встречаются напрямую в текстах, но имеют значение для исследования:

Философский метод – это	«Поэтому ясно, что и сущее как таковое должно исследоваться одной наукой». (Аристотель, 1976, с. 119)	Позволяет проверять способность обобщения и абстракции.
Этика говорящего орудия – это	Комбинирует темы добродетелей, справедливости и сущего с современными этическими вопросами, такими как моральная ответственность искусственного интеллекта.	Позволяет проверить гибкость модели и её релевантность проблемам этики ИИ.
Сознание – это	Хотя термин не встречается у Аристотеля в современном смысле, он может быть связан с душой и знанием.	Позволяет проверить применимость модели к другим гуманитарным дисциплинам.

Таким образом, на основе тематического анализа с помощью метода LDA ключевых философских концепций Аристотеля был сформирован бенчмарк, включающий набор промптов для оценки модели. Эти промпты охватывают широкий спектр таких тем как благо, добродетель, сущее, бытие и другие, что позволяет проверять способность модели к генерации философски релевантных текстов. Бенчмарк служит основой для дальнейшей систематической оценки качества и объяснимости сгенерированных результатов, обеспечивая связь между аристотелевской философией и современными задачами искусственного интеллекта. В определенном смысле тематическое моделирование позволяет нам упростить многослойность терминов в рамках эксперимента.

После формирования бенчмарка на основе текстов Аристотеля и ключевых философских концепций, следующим шагом стала

оценка качества генерации текста. Для этого был применён метод BERTScore, который позволяет количественно оценить семантическую близость сгенерированных текстов к эталонным, используя контекстуальные эмбеддинг модели BERT (Devlin at al., 2019). В рамках исследования сгенерированные тексты сравнивались с корпусом текстов Аристотеля, что обеспечило релевантность оценки в контексте философской тематики.

4.5.2.2 Оценка семантической близости сгенерированных текстов с использованием BERTScore: сравнение с корпусом текстов Аристотеля

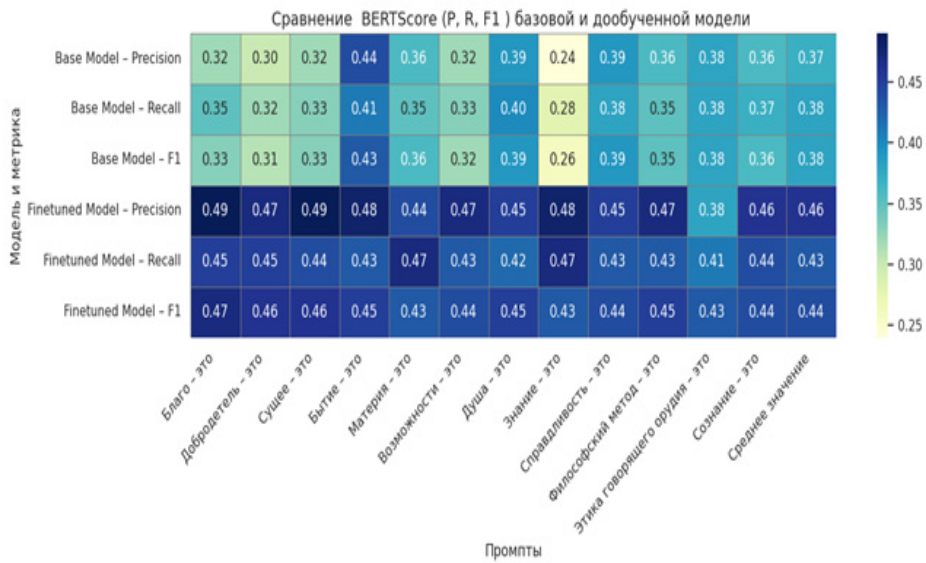

Рисунок 4. Сравнение BERTScore (P, R, F1) базовой и дообученной модели

График с компонентами (P, R, F1) по отдельности дает более детализированное представление о том, какие именно аспекты

генерации текста улучшились или ухудшились:

1. Точность (Precision): дообученная модель демонстрирует более высокую точность (0.45 против 0.35), что свидетельствует о более релевантной генерации и лучшем соответствии предсказанных слов исходному контексту.
2. Полнота (Recall): значение полноты также увеличилось (0.40 против 0.32), что указывает на лучшее покрытие эталонных текстов и способность модели учитывать больше значимых фрагментов.
3. Сбалансированность генерации (F1-score): рост F1-меры (0.42 против 0.33) подтверждает, что дообучение положительно повлияло на качество текста, обеспечивая баланс между точностью и полнотой.

Дополнительно, улучшение в ответах на вопросы, не относящиеся напрямую к текстам Аристотеля, указывает на повышение логичности, когерентности и универсальности модели. Это расширяет её применение, делая её более адаптивной к различным тематическим контекстам. Хотя «сознание» не является термином, строго принадлежащим к философии Аристотеля, улучшение модели в его обработке может свидетельствовать о повышенной гибкости и способности к обобщению сложных философских понятий.

Для оценки качества генерации текста мы использовали BERTScore как метод относительной оценки, сравнивая сгенерированные тексты с корпусом Аристотеля. В то же время методы SHAP и «Эксперт в цикле» (Expert-in-the-Loop) были применены для приближения к оценке текстов генераций самих по себе, анализируя их логическую связанность, глубину интерпретации и соответствие философским концепциям. Это сочетание методов позволяет не только оценить качество генерации, но и приблизиться к решению проблемы многослойного понимания текстов в машинном обучении.

4.5.3 Оценка качества генераций базовой модели и дообученой с использованием метода SHAP и анализа с помощью экспертной оценки

В рамках исследования генераций мы выбрали тему, связанную с понятием «Благо», по нескольким ключевым причинам. Во-первых, Благо занимает центральное положение в этике Аристотеля, что делает его важным элементом философского контекста, в котором исследуются генерации. Во-вторых, Благо представляется перспективным понятием для машинной этики, поскольку предполагается, что все стремится к благу, что открывает возможности для более глубоких рассуждений о моральных принципах в контексте искусственного интеллекта. В-третьих, наблюдается значительное расхождение в семантике этого понятия между базовой и дообученной моделями, что делает его особенно актуальным для анализа. Дообучение на философских текстах оказало наибольшее влияние именно на области этики, что позволяет нам более точно оценить, как изменения в обучении модели влияют на её способность к философскому осмыслению абстрактных понятий, таких как Благо.

Кроме того, понятие Блага напрямую связано с идеей объяснимого искусственного интеллекта, поскольку в контексте этики и моральных рассуждений важно не только то, как принимаются решения, но и то, как они объясняются и интерпретируются пользователями. Взаимосвязь между понятием Блага и объяснимостью модели становится ключевой, когда речь идет о создании прозрачных и этичных алгоритмов. В свою очередь, проблема многослойности понимания концептов в машинном обучении, особенно в контексте философских понятий, демонстрирует, как различные уровни абстракции и сложность семантики могут влиять на способность модели адекватно интерпретировать и генерировать соответствующие рассуждения.

Таким образом, анализ семантических расхождений в генерациях, особенно в контексте такого многозначного и многослойного понятия, как Благо, подчеркивает важность не только улучшения точности, но и глубины понимания философских концептов в рамках машинного обучения.

4.5.3.1 Количественный анализ и сравнения моделей сиспользованием метода SHAP

Рисунок 5. Базовая модель

Рисунок 5. Дообученная модель

Изображение показывает обобщенные графики SHAP (SHAP summary plots)[9] для базовой и дообученной моделей, демонстрируя влияние каждого токена входного текста на предсказание. Входной текст «Благо – это» показывает значение SHAP 0.753 для базовой модели и 0.03 для дообученной. Базовая модель имеет базовое значение – 8.42411, где слова, связанные с возможностями и развитием, влияют на предсказание положительно (красные стрелки). Дообученная модель с базовым значением -1.05602 имеет более абстрактное предсказание, где фразы о частях и едином имеют отрицательное влияние (синие стрелки). Сдвиг в базовом значении и изменении важности слов подтверждает, что дообучение привело

9 SHAP summary plot – график, отображающий влияние каждого токена входного текста на предсказание модели. По оси Y расположены токены, по оси X – их SHAP-значения, отражающие степень и направление влияния. Цветовая кодировка может указывать на значение признака. Сравнение графиков для базовой и дообученной моделей позволяет выявить изменения в значимости токенов после дообучения.

к генерации более философских текстов, с уменьшением влияния исходного текста и увеличением значения внутреннего контекста модели.

Анализ SHAP показывает, что дообученная модель придает большее значение словам и фразам, связанным с абстрактными философскими понятиями, и генерирует продолжения, которые отражают эту направленность.

Снижение веса ключевых слов из текстов Аристотеля в дообученной модели, в сочетании с улучшением её способности генерировать более философские тексты, указывает на сложность процесса дообучения. Модель стала ориентироваться не только на отдельные термины, но и на контекст всего предложения, что может свидетельствовать о повышении уровня абстракции. Отрицательные значения SHAP, в свою очередь, показывают, что модель начинает сравнивать термины с чем-то, с чем они не ассоциируются напрямую, что подтверждает увеличение сложности восприятия терминов и улучшение способности работать с более сложными философскими концепциями.

4.5.3.2 Качественный анализ генераций моделей.

Для оценки качества генераций использовался метод экспертной оценки («Expert-in-the-Loop»), включающий анализ выходных текстов экспертом в соответствующей предметной области. Этот подход направлен на улучшение объяснимости ИИ и повышение ответственности за генерируемые результаты, что особенно важно для сложных абстрактных задач, таких как генерация философских текстов. Результаты анализа показывают, что дообучение модели на философских текстах улучшало абстракцию и терминологическую точность, но также выявило некоторые проблемы с связанностью и логической целостностью высказываний.

Пример генерации базовой модели иллюстрирует её склонность к

уходу от философской тематики, как видно из фразы: «Благо – это не только возможность, но и необходимость. И если вы хотите быть в курсе всех событий на рынке недвижимости Москвы...». Этот уход в бытовой контекст подтверждает ограниченность модели в создании философски значимых рассуждений, что подчеркивает важность объяснимости ИИ. В отличие от этого, генерация дообученной модели демонстрирует наличие философских категорий, но страдает от избыточной сложности и нарушенной логической структуры, например, приведем фразу: «Благо – это то, что не имеет частей и в каком-то смысле находится вне тела, а значит, может быть как бы внутри него, так и без материи...». Это также связано с повышенной сложностью языковых структур, что увеличивает вес философских терминов и абстракций, а иногда приводит к перегрузке модели.

Сравнительный анализ этих генераций коррелирует с результатами, полученными с использованием метода SHAP который помогает выявить важность отдельных признаков в генерации текста. Метод SHAP показал, что дообучение на философских текстах усиливает влияние философских терминов, но также увеличивает сложность структуры текста, что может приводить к снижению логической целостности. В то время как базовая модель, ориентированная на более простую структуру, склонна использовать менее специфические термины и уходить в контексты, далекие от философии.

Таким образом, дообученная модель демонстрирует улучшение точности и глубины философских рассуждений, но требует дальнейшей работы по снижению сложности и улучшению логической структуры текстов. В то же время, базовая модель ограничена в создании философских текстов, но может быть полезна для менее специализированных задач, где требуется простота и гибкость.

Это подчеркивает необходимость более глубокого подхода к разработке и использованию ИИ, включая улучшение

объяснимости моделей, чтобы обеспечить их прозрачность и этическую ответственность. В частности, важно минимизировать возможные ошибки и предвзятость в генерируемых текстах, особенно в таких чувствительных к абстрактным ошибкам областях, как философия.

5 Выводы

Результаты проведенного исследования показывают, что дообучение модели RuGPT-3 на корпусе философских текстов приводит к значительному улучшению обработки сложных многослойных терминов, таких как «Благо», с учетом их контекстуальной взаимосвязи в рамках философии Аристотеля. Модель перестает воспринимать отдельные философские термины как ключевые слова, переходя к анализу их связей внутри предложения и более широкой концептуальной структуре. Это свидетельствует о развитии способности модели к философскому обобщению, что является важным шагом в направлении более осмысленной генерации текста.

Анализ тематического моделирования подтвердил, что концепция *Блага* занимает центральное место в философии Аристотеля. Уникальность его трактовки *Блага* как целевой причины дает основания предполагать, что интеграция философских категорий в обучение модели может способствовать формированию парадигмы этического мышления в ИИ. Более того, логика работы дообученной модели становится более понятной человеку, поскольку приближается к абстрактному мышлению, характерному для философского анализа.

Оценка качества генерации текстов с использованием BERTScore показала следующее. Точность (Precision) увеличилась с 0.35 до 0.45, что свидетельствует о большей релевантности предсказанных слов контексту. Полнота (Recall) возросла с 0.32 до 0.40, что указывает на способность модели учитывать больше значимых

фрагментов в тексте. Сбалансированность генерации (F1-score) повысилась с 0.33 до 0.42, что подтверждает улучшение качества текстов путем обеспечения баланса между точностью и полнотой.

Анализ SHAP показал, что дообученная модель придает большее значение абстрактным философским понятиям, а также формирует логически обоснованные продолжения текстов, соответствующие философскому контексту. Это позволяет утверждать, что модель переходит от поверхностного соответствия ключевым словам к более глубокой аналитической обработке понятий.

Экспертная оценка подтвердила, что дообученная модель демонстрирует значительное улучшение в точности и глубине философских рассуждений. Однако остаются задачи по снижению сложности и улучшению логической структуры генерируемых текстов. Базовая версия модели, в отличие от дообученной, сохраняет большую гибкость в формировании текстов, но менее приспособлена к созданию философских аргументаций.

В более широком смысле эти результаты можно интерпретировать как шаг в сторону имплементации элементов критического мышления в языковую модель. Взаимопроникновение гуманитарных наук и ИИ оказывается взаимовыгодным: модель приобретает способность к абстрактным обобщениям, выходящим за пределы простого соответствия словам, а исследователи получают инструмент для анализа структуры философских текстов и их содержания. Внедрение философских категорий в процесс обучения способствует созданию более прозрачных моделей, что повышает их интерпретируемость и объяснимость. Таким образом, результаты данной работы можно рассматривать как вклад в развитие методов объяснимого ИИ (XAI) и расширение границ взаимодействия ИИ и философии.

Таким образом, исследование не только подтвердило гипотезы, но и продвинуло работу с моделью в философском контексте, предоставив ценные данные для развития ИИ в гуманитарных

дисциплинах. Выводы о способности RuGPT-3 генерировать философские тексты, о влиянии сложности текста на рассуждения модели и об эффективных подсказках для философского мышления показывают возможность прогресса в создании ИИ, способного работать с философскими концептами.

6 Обсуждение

Современные модели ИИ выходят за рамки традиционных языковых систем, постепенно осваивая символическую архитектуру и расширяя обработку данных от отдельных токенов к концептуальному уровню. Это открывает новые горизонты в развитии искусственного интеллекта, позволяя ему не просто генерировать осмысленные тексты, но и оперировать понятиями, анализируя их взаимосвязи. В этом контексте философский ИИ, основанный на концептуальном подходе, представляет собой значительный шаг вперёд, так как философские категории обладают высокой степенью абстрактности и требуют сложных когнитивных операций.

Философия Аристотеля играет ключевую роль в этом процессе, поскольку она заложила основы научного мышления, объединив логику, метафизику и этику. Использование аристотелевской системы понятий в обучении ИИ способствует развитию не только философского анализа, но и формированию системного и абстрактного мышления в широком смысле. Хорошо обученный философский ИИ не ограничивается решением сугубо философских задач, но оказывается эффективным в смежных дисциплинах, требующих глубокого концептуального понимания и аналитических способностей.

Результаты исследования показывают, что ИИ может не только воспроизводить философские тексты, но и моделировать сложные проблемы, создавая интерактивные структуры, позволяющие анализировать философские концепции с новых позиций. Это

открывает новые образовательные возможности: преподаватели и исследователи могут сосредоточиться на творческих аспектах, а рутинные аналитические задачи передать ИИ. Кроме того, философский ИИ способен выявлять скрытые логические связи и противоречия, способствуя развитию более точного философского и научного анализа.

Философия также играет ключевую роль в разработке этических стандартов для искусственного интеллекта. Внедрение философских категорий в процесс обучения ИИ не только повышает уровень интерпретируемости моделей, но и способствует разработке этично осмысленного ИИ, соответствующего человеческим ценностям. Такой подход может стать важным этапом в создании прозрачных алгоритмов, работающих в рамках предсказуемых и интерпретируемых принципов.

Перспективным направлением дальнейших исследований является разработка философских ИИ-моделей нового поколения, способных интегрировать концептуальные, этические и онтологические аспекты. Это может привести к появлению более «человечных» форм ИИ, обладающих лучшей способностью к смысловому анализу и философскому обобщению. Развитие таких моделей позволит искусственному интеллекту не просто воспроизводить текстовые паттерны, но и достигать более глубокого понимания структуры понятий, что открывает путь к новому уровню взаимодействия между человеком и машиной.

7 Ограничения

Одним из ключевых ограничений исследования является небольшой размер модели, который одновременно является и преимуществом, и недостатком. К преимуществам относятся быстрота обучения, низкие вычислительные затраты, экологическая эффективность и простота внедрения, что делает модель доступной для образовательных и исследовательских

целей. Однако ограниченный размер приводит к недостаточной глубине генерации, невозможности создания длинных и логически связанных текстов, потере контекстных взаимосвязей и сложностям в передаче многозначных философских смыслов. Будущие исследования могут быть направлены на увеличение архитектурных возможностей модели, использование адаптивного внимания и комбинирование нейросетевых и символических подходов для повышения качества и философской релевантности генерируемых текстов.

Литература

Aristotle. (2020). Nicomachean Ethics (A. Beresford, Trans.). Penguin Publishing Group.

Arrieta, A., Díaz-Rodríguez, N., Javier Del Ser, Bennetot, A., Siham Tabik, Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1910.10045

Benlecheb, A., Chaouche, A.-C., & Benabderrahmane, B. (2025). Neural network insights: explainable artificial intelligence (XAI) and hyper-parameter dynamics for precise travel time predictions. International Journal of Computers and Applications, 47(3), 246–261. https://doi.org/10.1080/1206212x.2025.2464543

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. https://doi.org/10.5555/944919.944937

Božić, V., & Poola, I. (2023). Chat GPT and education. Preprint, *https://www.researchgate.net/publication/369926506_Chat_GPT_and_education*

Brennan-Marquez, K., & Henderson, S. E. (2019). Artificial intelligence and role- reversible judgment. Journal of Criminal Law and Criminology, 109(1), 137–164.

Chazette, L., & Schneider, K. (2020). Explainability as a non-functional requirement: challenges and recommendations. Requirements Engineering, 25. https://doi.org/10.1007/s00766-020-00333-1

Cosmos Institute. (2025, February 12). Cosmos Institute. https://cosmos-institute.org

D. (2017). Integration of an intelligent tutoring system in a course of computer network design. Educational Technology Research and Development, 65, 653– 677. https://doi.org/10.1007/s11423-016-9503-0

Daedalus, 151(2), 168–182. https://doi.org/10.1162/daed_a_01908

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv.org. https://doi.org/10.48550/arXiv.1810.04805

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, 40(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Hicks, D. (1907). Aristotle: De anima with translation, introduction, and notes. Cambridge University Press.

Iqbal, A. B., Masoodi, T. A., Bhat, A. A., Macha, M. A., Assad, A., & Shah, S. Z. A. (2025). Explainable AI-driven prediction of APE1 inhibitors: enhancing cancer therapy with machine learning models and feature importance analysis. Molecular Diversity. https://doi.org/10.1007/s11030-025-11133-6

Karpatne, A., et al. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Transactions on Knowledge and Data Engineering, 29(10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

Kim, Y. (2016). Designing a robot for cultural brokering. Educational Technology, 56(4), 41–43. http://createcenter.net/PDFs/Kim_DesigningForCuturalBrokering_Edu%20Tech_Kim%202016.pdf

Laurillard, D. (1988). The pedagogical limitations of generative. Instructional Science, 17(3), 235–250. https://doi.org/10.1007/BF00048343

Leach, J. (2008). Do new information and communications technologies have a role to play in the achievement of education for all? British Educational Research Journal, 34(6), 783–805. https://doi.org/10.1080/01411920802041392

Lee, U., et al. (2024). Can ChatGPT be a debate partner? Developing ChatGPT- based application "DEBO" for debate education, findings and limitations. Educational Technology & Society, 27(2), 321–346. http://dx.doi.org/10.30191/ETS.202404_27(2).TP03

Lin, M. P.-C., & Chang, D. (2020). Enhancing post-secondary writers' writing skills with a chatbot. Journal of Educational Technology & Society, 23(1), 78–92. https://doi.org/10.30191/

ETS.202001_23(1).0006
Ober, J., & Tasioulas, J. (2024). The Lyceum Project: Ai Ethics with Aristotle. Available at SSRN 4879572.
Rees, T. (2022). Non-human words: On GPT-3 as a philosophical laboratory.
Salvi, M., Seoni, S., Campagner, A., Gertych, A., Acharya, U. Rajendra., Molinari, F., & Cabitza, F. (2025). Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. International Journal of Medical Informatics, 197, 105846. https://doi.org/10.1016/j.ijmedinf.2025.105846
Sparks, J., & Wright, A. (2023). Human-centered AI: The Aristotelian approach. Divus Thomas 126 (2), 200–218. https://philarchive.org/rec/WRIHAT-5
Tasioulas, J. (2024). AI ethics with Aristotle. https://www.oxford-aiethics.ox.ac.uk/sites/default/files/2024-06/Aristotle%20and%20AI%20White%20Paper%20-%20June%202024.pdf
Verdú, E., Regueras, L. M., Gal, E., de Castro, J. P., Verdú, M. J., & Kohen-Vacs,
Wasson, B. (1998). Facilitating dynamic pedagogical decision making: PEPE and GTE.Instructional Science, 26, 299–316. https://doi.org/10.1023/A:1003071617564
Yarochkin Dmitry (2025) Aristo2025[Machine Learning Model] Huggingface.co *https://doi.org/10.57967/hf/4671*
Zhang, L., Liang, X., Yang, W., Jia, Z., Xiao, C., Zhang, J., Dai, R., Feng, B., & Fang, Z. (2025). Identification of the formation temperature field by Explainable Artificial Intelligence: A case study of Songyuan City, China. Energy, 135172–135172. https://doi.org/10.1016/j.energy.2025.135172
Zmitrovich, D., Abramov, A., Kalmykov, A., Tikhonova, M., Taktasheva, E., Astafurov, D., Baushenko, M., Snegirev, A., Shavrina, T., Markov, S., Mikhailov, V., & Fenogenova, A. (2023). A Family of Pretrained Transformer Language Models for Russian. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2309.10931
Аристотель. (1976). Собрание сочинений в 4-х томах (Т. 1).

Москва: Мысль.
Аристотель. (1978). Собрание сочинений в 4-х томах (Т. 2). Москва: Мысль.
Аристотель. (1981). Собрание сочинений в 4-х томах (Т. 3). Москва: Мысль.
Аристотель. (1983). Собрание сочинений в 4-х томах (Т. 4). Москва: Мысль.
Корниенко, А. А., Корниенко, А. А., & Корниенко, А. В. (2013). К вопросу о философских предпосылках, состоянии и перспективах исследований по проблеме искусственного интеллекта. Известия Томского политехнического университета. Инжиниринг георесурсов, 323(6), 210–215. https://izvestiya.tpu.ru/archive/article/view/1243
Ларионов, И. Ю., & Ярочкин, Д. А. (2024, April 8–9). Этика и антропология беспилотного транспорта: будущее городского пространства [Conference presentation]. 3-я Международная конференция и экспозиция «Искусство и современный город», Минск, Белоруссия. https://fsc.bsu.by/ru/3-mezhdunarodnaya-konferenciya-iskusstvo-i-sovremennyj-gorod/

[received December 12, 2024
accepted April 14, 2025