

УДК 004.032.26
doi:10.18720/SPBPU/2/id24-515

*Доброходов Ярослав Иванович*¹,
ассистент;
*Сурина Алла Валентиновна*²,
доцент, канд. тех. наук, доцент

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ: ПРОГНОЗИРОВАНИЕ РАЗВИТИЯ ПО МЕТОДУ ГАРТНЕРА

^{1,2} Россия, Санкт-Петербург,
Санкт-Петербургский политехнический университет Петра Великого,
¹ dobroh_yai@spbstu.ru, ² surina_av@spbstu.ru

Аннотация. В статье представлен составленный по методу Гартнера прогноз развития нейронных сетей с применением больших языковых моделей в ближайшие пять лет. Особое внимание уделено вопросам взаимодействия с конечным потребителем и роли нейронных сетей в современном информационном пространстве.

Ключевые слова: искусственный интеллект, большие языковые модели, нейронные сети, технологическое прогнозирование, цикл зрелости технологий, кривая Гартнера, когнитивное моделирование.

*Yaroslav I. Dobrokhodov*¹,

Assistant lecturer;

*Alla V. Surina*²,

Associate Professor, Candidate of Technical Sciences

LARGE LANGUAGE MODELS: DEVELOPMENT FORECASTING USING THE GARTNER METHOD

^{1,2} Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia,

¹ dobroh_yai@spbstu.ru, ² surina_av@spbstu.ru

Abstract. The article presents a forecast for the development of neural networks using large language models in the next five years, compiled using the Gartner method. Particular attention is paid to issues of interaction with the end consumer and the role of neural networks in the modern information space.

Keywords: artificial intelligence, large language models, neural networks, technological forecasting, technology maturity cycle, Gartner hype cycle, cognitive modeling.

Введение

В последние два года наблюдается устойчивый рост интереса к технологиям, использующим искусственный интеллект. Одним из катализаторов данного процесса стали нейронные сети. Данная статья является попыткой понять актуальное состояние технологии с применением больших языковых моделей и попытаться спрогнозировать дальнейшее направление ее развития. Это помогло бы не только понять, какая часть окружающей ее информации из различных открытых источников является достоверной и заслуживающей внимания, но и получить представление о том, какие новые вызовы и возможности она может принести с собой в ближайшей перспективе.

Основная часть

В качестве основного метода для построения прогноза была выбрана кривая Гартнера, наиболее точно представляющая жизненный цикл технологии в терминах ее зрелости. Данный метод основывается на утверждении, что любая новая технология проходит через пять фаз, отражающих колебания возлагаемых на нее ожиданий [4]. Предполагается, что эти колебания систематически совпадают с прохождением различных этапов разработки технологии. Сама модель связана с коммерческой стороной новой технологии и отражает существующую в информационном обществе связь между выводом разработки на рынок и формированием общественного мнения о ней [1].

Использованная для построения прогноза информация относится к трем категориям: технические данные, поведение целевой аудитории и

коммерческая составляющая. Технические данные, полученные в процессе ознакомления с находящимися в свободном доступе наработками, использовались для оценки того, насколько информация из двух других категорий близка к реальному состоянию продукта. Поведение целевой аудитории, оценка которого осуществлялась на основе отзывов и обсуждений, является ключевым фактором для построения кривой ожиданий, а также помогает выявить ряд уникальных проблем, связанных сприятием искусственного интеллекта массовым потребителем. Третья категория включает в себя данные о том, насколько успешно функционируют ведущие разработку продуктов с применением нейронных сетей компании, а также об общем состоянии рынка высоких технологий. Выводы по этой категории основывались на изменениях рыночных котировок, объема инвестиций, количества рабочих мест и ценовой политики.

Для определения и корректировки предположительных сроков прохождения тех этапов развития, которые рассматриваются в статье в качестве предстоящих, был проведен сравнительный анализ жизненного цикла иных разработок с применением генеративного искусственного интеллекта и их предшественников, в результате которого также было выявлено несколько общих для всей сферы проблем, которые предстоит решить в будущем.

Графически полученный прогноз представлен кривой зависимости возлагаемых на технологию пользовательских ожиданий от времени (см. рис. 1) и сопровождается таблицей с расшифровкой пронумерованных этапов (см. табл. 1).

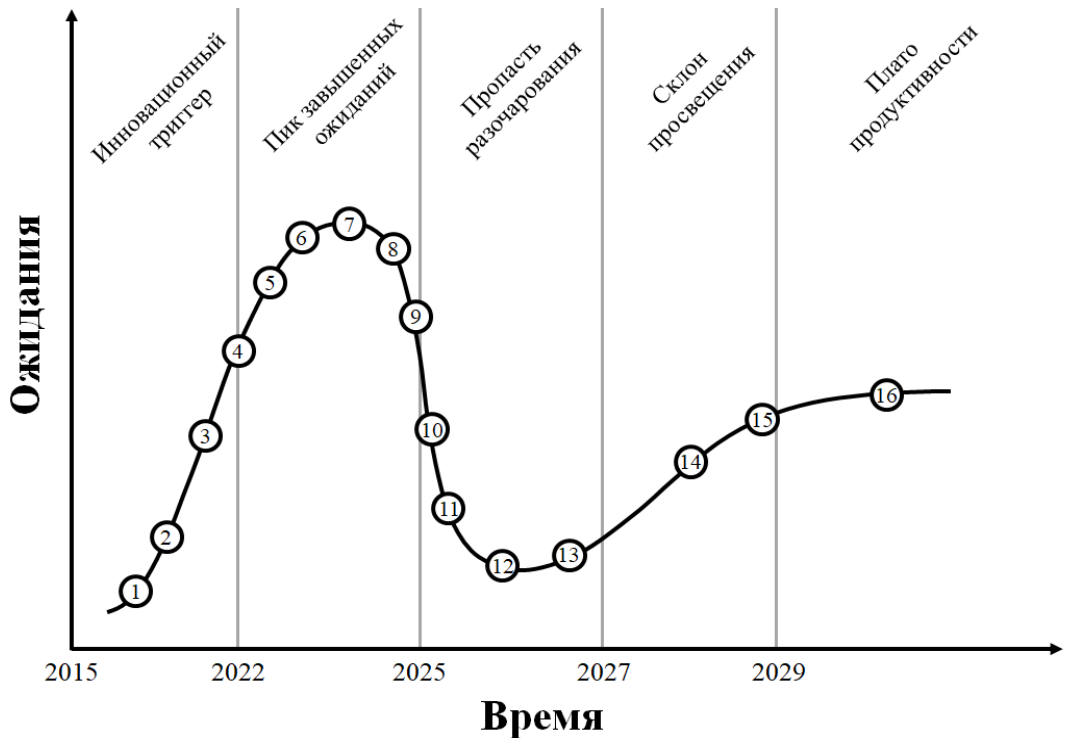


Рис. 1. Кривая Гартнера

Таблица 1

Этапы цикла зрелости

Фаза	№	Этап	Примерная дата начала этапа
Инновационный триггер	1	Начало разработки технологии	Январь 2016
	2	Первый этап финансирования	Апрель 2016
	3	Первое поколение продукта	Ноябрь 2022
	4	Рост внимания целевой аудитории	Декабрь 2022
Пик завышенных ожиданий	5	Начало широкого освещения в СМИ	Март 2023
	6	Появление новых поставщиков	Май 2023
	7	Усиление активности аудитории	Июнь 2023
	8	Первые негативные отзывы в СМИ	Август 2023
	9	Консолидация поставщиков	Сентябрь 2023
Пропась разочарования	10	Первые регулирующие технологию законы	Весна 2025
	11	Второй раунд финансирования	Лето 2025
	12	Технологию приняло менее 5 % аудитории	Конец 2025 – начало 2026
	13	Второе поколение продукта	Весна – лето 2027
Склон просвещения	14	Формализация успешного опыта	Осень 2027
	15	Третье поколение продукта	2028
Плато продуктивности	16	Интенсивный рост использования	После 2028

Этап инновационного триггера в данном случае начинается в 2015 году, когда на территории США основывается некоммерческая организация OpenAI, которой удастся быстро привлечь инвесторов и начать разработку серии проектов, в основу одного из которых ложатся большие языковые модели. Три года спустя появляется первое поколение генеративных предобученных трансформеров или GPT, больших языковых моделей, обученных генерировать текст, схожий с человеческим. Еще через четыре года, в ноябре 2022 года, представлено первое поколение чат-ботов с генеративным искусственным интеллектом на основе GPT третьей модели. Машина, способная создавать уникальный персонализированный текст в

процессе общения с человеком, быстро завоевывает популярность в сети и привлекает внимание СМИ. Позитивная реакция общественности помогает коммерциализировать разработку, что, в свою очередь, начинает формировать облик будущего рынка. К концу 2023 года это создает условия, позволяющие появиться альтернативным поставщикам, предлагающим свои решения на основе общедоступного ядра технологии.

Несмотря на то, что пик завышенных ожиданий в полной мере оправдывает свое название, полученная от пользователей обратная связь в данный период времени позволяет составить первые представления о том, в каком направлении будет развиваться технология, чтобы отвечать внешнему спросу, и какие препятствия ей для этого придется преодолеть.

Также в этот период можно выделить основные источники беспокойства потребителей. Поскольку средний пользователь еще не в полной мере понимает реальные пределы возможностей нейросетей на данном этапе разработки, его воображение рисует гипертрофированные, но от того более различимые картины потенциальных кризисов, связанных с новой технологией.

Многие традиционно боятся, что машины могут заменить их в ближайшем будущем, оставив без работы. Действительно, нейросети начинают, например, активно использоваться в креативных сферах и неизбежно займут какой-то процент рынка, но вместо того, чтобы вытеснить с него всех прочих работников, искусственный интеллект просто станет одним из них. Уже сейчас можно с относительной легкостью опознать изображения, аудио, тексты и даже программный код, созданные нейросетями, поскольку те обучены придерживаться определенного воспроизводимого паттерна. Переобучение или создание новой специализированной модели требует времени и денег. Сам период обучения может расти или сокращаться в зависимости от того, как эта сфера будет регулироваться в будущем, но полностью от него избавиться не удастся.

Еще один классический страх – неуправляемый искусственный интеллект с неограниченной властью. Пожалуй, эта тема еще достаточно долгое время будет оставаться в рамках научной фантастики. Современная нейросеть – это специализированный инструмент с ограниченным набором функций. При желании ее можно научить серии трюков, которые она раз за разом вполне успешно будет повторять, но только при воздействии внешнего импульса. Сама система, по которой строятся нейросети, остается имитацией ограниченного набора простейших логических принципов, выделенных человеком.

Третий источник беспокойства представляется новым и куда более интересным. Рост значимости информации в современном обществе привел к тому, что вопросами авторского права в той или иной степени сейчас озабочен каждый, кто хоть раз писал что-то в сети. Любую модель нужно

обучать на наборе исходных данных, а современные модели обучаются на огромных архивах информации, многие из которых находятся в свободном доступе, но некоторые остаются закрытыми [3]. Ожидаемо, со временем и под влиянием имевших место несколько лет назад крупных скандалов, никак не связанных с рассматриваемой технологией, в сознании некоторой части пользователей это привело к устойчивой идее, что нейросети не только нарушают авторские права, но даже «крадут» чужие работы. В настоящее время вопрос соблюдения авторского права довлеет над всей сферой искусственного интеллекта, из-за чего правительства ряда стран независимо друг от друга обсуждают вопросы законодательного регулирования, призванного разрешить эту проблему.

Однако пользовательский интерес преодолел свой пик в первой половине 2023 года, а затем начал снижаться. Помимо того, что за внимание пользователей непрерывно идет борьба, в которую регулярно вступают все новые инфоповоды, у происходящего в данный момент спада интереса и доверия к нейросетям были и внутренние причины. При этом важно отметить, что основное влияние оказали не упомянутые выше страхи пользователей, а их обманутые ожидания. Эффект новизны в сочетании с отсутствием технологической компетентности рядового пользователя привел к тому, что многие из них безоговорочно начали доверять технологии, чей основной задачей в настоящее время является необходимость как можно убедительнее имитировать человеческую речь, а не предоставлять абсолютно точную и фактически верную информацию [2].

С течением времени это привело к накоплению значительной массы провалов, многие из которых оказались достаточно серьезными. Так, например, в мае 2023 года один американский юрист в рамках судебного процесса воспользовался обзором предыдущих дел со схожим кейсом, информацию по которым он ранее запросил у чат-бота OpenAI. Сторона ответчика вскоре обнаружила, что шести из представленных дел никогда не существовало, и последующее расследование это подтвердило. Позднее юрист под присягой утверждал, что также запрашивал у чат-бота подтверждение подлинности предоставленной информации, и тот раз за разом утверждал, что все приведенные дела были реальными.

Описанная ситуация возникла из комбинации двух факторов, человеческого и чисто технического. После того, как подобная разработка попадает в открытый доступ, ее среднестатистическим пользователем с высокой вероятностью становится человек в действительности к работе с ней не подготовленный. Работа такого пользователя с нейросетью будет сопровождаться целой серией проблем, вызванных непониманием правил, по которым она функционирует, и прежде всего от этого пострадает качество формируемых запросов. А поскольку нейросеть работает с большими

данными, возможности пользователя проверить достоверность получаемой информации будут существенно ограничены [6].

Тем не менее, такая разработка имеет право быть представленной широкому кругу неквалифицированных пользователей, поскольку для них она позиционируется, прежде всего, как средство развлечения, а они, в свою очередь, способствуют ее тестированию и дальнейшему обучению. Судя по всему, убеждение в том, что данная разработка является готовым продуктом, способным выполнять функции виртуального помощника и предоставлять объективно верные данные, сформировалась непосредственно в среде конечных пользователей. Причиной этого, скорее всего, послужила техническая особенность подобных разработок, на которую большинство пользователей предпочитает не обращать внимания.

В действительности главной текущей задачей разработки больших языковых моделей является достижение наибольшего соответствия получаемых текстов тому, что мог бы написать живой человек. Проще говоря, нейросети, скорее всего, научатся сообщать то, что нам хотелось бы услышать раньше, чем приводить фактически верную информацию. В результате подобной тенденции генерируемый текст для многих пользователей звучит достаточно убедительно, чтобы пренебречь дополнительной проверкой результатов.

Здесь также нужно обратить внимание на то, что GPT является коммерческим продуктом и последняя на данный момент версия GPT-4 доступна исключительно в рамках платной подписки. Очевидно, что для продвижения своей продукции OpenAI необходимо смещать акцент на сильные стороны и успехи разрабатываемой модели, из-за чего со временем у массового потребителя продолжает укрепляться и развиваться свое видение максимально инновационной и эффективной технологии, отличной от того, что, например, известно работавшим с моделью программистам.

Растущее разочарование в нейросетях совпало с намечающейся тенденцией к уходу инвесторов с рынка высоких технологий. В период пандемии внимание инвесторов привлекла сначала сфера цифровых развлечений, а затем и остальная часть цифрового рынка, что привело к притоку капитала и бурному развитию рынка, однако текущие результаты этого развития, похоже, оказались неудовлетворительными, и поставщики вынуждены искать способы минимизировать убытки. Сложившаяся на конец 2023 года ситуация инициировала процесс консолидации, когда более устойчивые компании начали поглощать небольшие фирмы и стартапы, не сумевшие адаптироваться к изменившимся условиям.

Еще одним из способов возместить убытки стал постепенный отказ некоторых поставщиков от политики региональных цен, что ведет к росту стоимости технологических решений от крупнейших, в частности располагающихся на территории США поставщиков в других уголках мира. В

долгосрочной перспективе это может позитивно повлиять на развитие нейросетей с языковыми моделями.

В настоящий момент эта разработка удерживается в занимаемой ею нише гораздо увереннее, чем другие имевшие место в последние десять лет попытки создания инновационных ИТ-продуктов, оказавшиеся провальными. Нейросети явно останутся с нами надолго, поскольку успели прочно войти в сферу заинтересованности ИТ-специалистов, которые будут поддерживать умеренный, но стабильный интерес к данной теме.

Продолжающееся накопление опыта вкупе с кажущимся неизбежным ростом стоимости уже представленных на рынке решений создают благоприятную почву для развития собственных разработок в других странах, обладающих нужным научно-техническим потенциалом. Период затишья и спада интереса со стороны рядового пользователя может быть использован как окно для сокращения отставания и подготовки собственного пула квалифицированных специалистов, что позволит подключиться к процессу дальнейшего развития технологии к тому моменту, когда появление второго поколения нейросетевых сервисов станет вопросом обозримого будущего.

Однако одним из камней преткновения по-прежнему остаются вопросы регулирования развития искусственного интеллекта и охраны авторского права, которые необходимо будет решить, чтобы быстрее вернуть доверие инвесторов и потенциальных клиентов [7]. Помимо этого, для создания второго поколения потребуется решить вопрос достоверности генерируемой нейросетью информации. Если эта проблема будет преодолена, то перед новыми моделями откроется перспектива интеграции в более сложные и функциональные системы с более высоким уровнем ответственности.

К тому моменту, когда ситуация в сфере стабилизируется, и после решения ряда сохраняющихся технических проблем будет сформирована универсальная методика, нейронные сети начнут постепенно догонять ранее возложенные на них ожидания. Дальнейший качественный прорыв, который позволит технологии выйти на стабильный рост использования, скорее всего, будет сопряжен с параллельным развитием иных разработок в области человеко-машинных интерфейсов и автоматизированных устройств, что позволит активнее применять обучающиеся модели за пределами цифрового пространства.

Таким образом, несмотря на продолжающееся в настоящий момент падение интереса со стороны среднестатистического потребителя, растущее внимание в ряде стран, не включившихся в разработку на ранних этапах, позволит продолжать развитие технологии [5]. Это создает в ближайшие несколько лет благоприятное окно для подготовки специалистов, что также поспособствует решению ряда существующих технических задач.

Заключение

Описанная динамика в ближайшее время породит определенную дискуссию и кратковременный повторный рост интереса в различных регионах, как к самой технологии, так и к вопросу поиска альтернативных решений для ее вывода на рынок. Тема может представлять интерес для задействованных в сфере информационных технологий разработчиков и управляющих, а также преподавателей высших учебных заведений, осуществляющих подготовку специалистов соответствующих направлений, поскольку подобных обсуждений неизбежно стоит ожидать и в отечественном информационном пространстве, что делает ближайший год наиболее благоприятным для разработки соответствующих образовательных программ и привлечения широкой аудитории. Однако развитие дискуссии в русскоязычном пространстве, повторяя глобальный процесс, также со значительной вероятностью будет сопровождаться завышенными ожиданиями и связанными с ними ложными толкованиями реальных возможностей и характеристик нейронных сетей.

Список литературы

1. Dedehayir O., Steinert M. The hype cycle model: A review and future directions [Electronic resource] // Technological Forecasting and Social Change. – 2016. – No. 108. – Pp. 28–41. – DOI:<https://doi.org/10.1016/j.techfore.2016.04.005>. – URL: <https://www.sciencedirect.com/science/article/abs/pii/S0040162516300270> (access date: 31.05.2024).
2. Donges N. 4 Disadvantages of neural networks [Electronic resource] // National Techs & Startups, Built In. – URL: <https://builtin.com/data-science/disadvantages-neural-networks> (access date: 31.05.2024).
3. Guadamuz A. Artificial intelligence and copyright [Electronic resource] / World Intellectual Property Organization (WIPO). – URL: https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html (access date: 31.05.2024).
4. Perri L. Gartner hype cycle research methodology [Electronic resource] // Gartner Inc. – URL: <https://www.gartner.com/en/research/methodologies/gartner-hype-cycle> (access date: 31.05.2024).
5. Воробьева В. А. Развитие искусственного интеллекта [Электронный ресурс] // Сайт министерства экономического развития Российской Федерации. – URL: https://www.economy.gov.ru/material/departments/d01/razvitie_iskusstvennogo_intellekta/ (дата обращения: 31.05.2024).
6. Иванько А. Ф., Иванько М. А., Сизова Ю. А. Нейронные сети: общие технологические характеристики [Электронный ресурс] // Научное обозрение. Технические науки. – 2019. – № 2. – С. 17–23. – URL: <https://science-engineering.ru/ru/article/view?id=1236> (дата обращения: 31.05.2024).
7. Стенюшкин А. А. Развитие нейросетей в России: возможности и угрозы [Электронный ресурс] // Новости цифровой трансформации, телекоммуникаций, вещания и ИТ. – URL: <https://www.comnews.ru/content/233307/2024-05-23/2024-w21/1013/razvitie-neyrosetey-rossii-vozmozhnosti-i-ugrozy> (дата обращения: 31.05.2024).