УДК 004.8

РАЗВИТИЕ И ПРИМЕНЕНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Семерня А.С., магистрант гр. ПИм-231, II курс,
Кузбасский государственный технический университет
имени Т.Ф. Горбачева, г. Кемерово

Введение

Большие языковые модели (LLM, Large Language Models) стали одним из ключевых прорывов в области искусственного интеллекта. Это нейронные сети с миллиардами параметров, обученные на огромных массивах текстовых данных, способные генерировать, анализировать и преобразовывать человеческий язык [1-4]. Их развитие началось с архитектуры Transformer, представленной Google в 2017 году, и достигло пика с появлением ChatGPT и GPT-439. Сегодня LLM применяются в чат-ботах, переводчиках, системах анализа данных и даже медицине, демонстрируя потенциал для трансформации множества отраслей[2, 5].

Суть LLM

LLM – это системы глубокого обучения, основанные на архитектуре Transformer. Они обрабатывают текст как последовательность токенов (слов или их частей) и учатся предсказывать наиболее вероятное продолжение фразы, анализируя контекст[5, 7]. Ключевые этапы их работы:

1) предобучение [6] – модель изучает общие закономерности языка на огромных корпусах текста (например, книгах, статьях, веб-страницах);

2) тонкая настройка [1, 8] – адаптация под конкретные задачи (например, медицинскую диагностику или программирование) с использованием узкоспециализированных данных;

3) Генерация ответа — на основе механизма внимания (attention), который определяет значимость каждого слова в контексте[7]

Архитектура Transformer включает энкодеры и декодеры, что позволяет обрабатывать текст параллельно, а не последовательно, значительно ускоряя обучение[5, 7].

Transformer: основа современных LLM

Архитектура Transformer [9], стала революцией в обработке естественного языка (NLP). Её ключевые компоненты:

- Механизм внимания (Self-Attention): Позволяет модели анализировать связи между словами в предложении, независимо от их позиции. Например, в фразе «The cat drank milk because it was thirsty» механизм определяет, что «it» относится к «cat», а не к «milk».

- Позиционное кодирование (Positional Encoding): Добавляет информацию о порядке слов, так как Transformer не имеет встроенной последовательной обработки (в отличие от RNN).
- Энкодеры и декодеры:
  - энкодер преобразует входной текст в контекстные представления;
  - декодер генерирует выходной текст, используя эти представления.

Популярные модели и их применение

- GPT-4 (OpenAI). Особенности: 1.7 трлн параметров, мультимодальность (работа с текстом и изображениями) [3, 8]. Применение: ChatGPT для диалогов, генерация кода, создание контента [5, 8].
- BERT (Google). Особенности: двунаправленный анализ контекста, 340 млн параметров [3, 8]. Применение: улучшение поисковых систем, классификация текста [8].
- GigaChat (Сбер). Особенности: поддержка русского языка, интеграция с Kandinsky для генерации изображений [1]. Применение: SEO-оптимизация текстов, анализ финансовых документов [1, 4].
- LLaMA 2 (Meta). Особенности: открытые веса, эффективность на малых вычислительных ресурсах [3, 8]. Применение: исследования в области ИИ, локальные решения для бизнеса [8].
- Claude 3.5 (Anthropic). Особенности: обработка длинных контекстов (до 1 млн токенов), анализ видео [8]. Применение: юридический анализ, генерация технической документации.

Ограничения и вызовы

- Ресурсоемкость: обучение GPT-3 требует энергии, сопоставимой с годовым потреблением 120 домохозяйств [8].
- Предвзятость: модели могут воспроизводить стереотипы из обучающих данных [2, 8].
- «Галлюцинации»: генерация ложной информации из-за статистического подхода [3, 6].

Заключение

Большие языковые модели открывают новые горизонты в автоматизации рутинных задач, образовании и креативных индустриях. Однако их развитие требует решения этических и технических проблем, таких как энергоэффективность и контроль за достоверностью данных [5, 8]. Будущее LLM связано с мультимодальностью (обработка текста, звука, видео) и созданием компактных версий для локального использования.

Список литературы:

1. Что такое LLM? Обзор больших языковых моделей и их применение. – URL: https://developers.sber.ru/help/gigachat-api/large-language-models (дата обращения: 20.03.2025).

2. Большие языковые модели: что это такое и как они работают. – URL: https://just-ai.com/blog/bolshie-yazykovye-modeli-chto-eto-takoe-i-kak-oni-rabotayut (дата обращения: 31.03.2025).

3. LLM: как работают языковые модели для чат-ботов и умных поисковиков. – URL: https://trends.rbc.ru/trends/industry/6784cece9a7947485ec2f599 (дата обращения: 20.03.2025).

4. Обзор небольших больших языковых генеративных моделей: GPT и русские версии. – URL: https://sber.pro/publication/nebolshie-bolshie-yazikovie-modeli-kakie-prikladnie-zadachi-oni-reshayut/ (дата обращения: 26.03.2025).

5. Что такое большие языковые модели?. – URL: https://aws.amazon.com/ru/what-is/large-language-model/ (дата обращения: 29.03.2025).

6. Большие языковые модели и их особенности. – URL: https://na-journal.ru/12-2023-informacionnye-tekhnologii/7492-bolshie-yazykovye-modeli-i-ih-osobennosti (дата обращения: 26.03.2025).

7. Гайд: как создать большие языковые модели. – URL: https://scand.com/ru/company/blog/building-and-training-large-language-models-your-ultimate-guide/ (дата обращения: 29.03.2025).

8. Как работают языковые модели. – URL: https://habr.com/ru/articles/825690/ (дата обращения: 28.03.2025).

9. Attention Is All You Need. – URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (дата обращения: 26.03.2025).