Конспект лекции: Современные ИИ и загадки нейросетей

1 Введение

- **Ключевой тезис**: Нейросети — не «чёрные ящики», важно понимать их внутреннюю математику из-за ненадёжности работы
- **Основные направления лекции**:
  1. Эволюция подходов в ИИ
  2. Анализ двух современных исследований о странных особенностях нейросетей

2 История машинного перевода

2.1 Правиловый подход (1950-2000-е)

- Попытка формализации лингвистических правил:
  - Словари и грамматические конструкции
  - Учёт исключений и стилей речи
- **Проблемы**:
  - Высокие трудозатраты
  - Низкое качество (пример: мемы о плохих переводах)
  - Отсутствие масштабируемости

2.2 Data-driven подход (с 2000-х)

- Использование параллельных текстов (книги, субтитры)
- Техническая реализация:
  - Модели с миллиардами параметров
  - Оптимизация через методы машинного обучения
- **Результат**: Качество перевода значительно улучшилось (Google Translate, Яндекс.Переводчик)

3 Современные языковые модели (LLM)

3.1 Основные принципы

- Предсказание следующего слова → генерация осмысленного текста
- Два этапа обучения:
  1. На «сырых» данных (весь интернет)
  2. На верифицированных данных (коррекция токсичности)
- **Требования**:
  - Триллионы токенов для обучения
  - Кластеры GPU/TPU

3.2 Практические применения

- **Vibe-coding**:
  - GitHub Copilot и аналоги
  - Перспектива для не-программистов
- **AlphaGeometry**:
  - Решение задач через формальные доказательства
  - Язык для верифицируемых рассуждений
- **Агенты**:
  - Интеграция с внешними API
  - Пример: запрос погоды

4 Ключевые исследования

4.1 Феномен супервесов (Superweights)

- **Открытие**: Единичные параметры, критически влияющие на работу модели
- **Эффекты**:
  - Зануление → полный сбой модели
  - Увеличение → улучшение качества (парадокс)
- **Практическое значение**:
  - Проблемы с устойчивостью
  - Оптимизация квантизации

4.2 Хранение знаний в LLM

- **Локализация**: Полносвязные слои трансформеров
- **Эксперимент**:
  - Подмена входных данных
  - Восстановление отдельных блоков
- **Выводы**:
  - Возможность точечного редактирования фактов
  - Угрозы безопасности (скрытые модификации)

5 Выводы и перспективы

- **Текущие ограничения**:
  - Исчерпание интернет-данных
  - Экспоненциальный рост стоимости обучения
- **Перспективные направления**:
  - Мультимодальность (видео, аудио)
  - Генерация проверяемых данных (математика, код)
  - Архитектуры Mixture of Experts
- **Ключевой вывод**: Необходим баланс между мощностью моделей и их интерпретируемостью