300 ЛЕТ СПбГУ

# Система прогнозирования событий с использованием RAG

**Олизько Степан Сергеевич**
Студент группы 22.Б15-пу

**Антон Юрьевич Першин**
Научный руководитель

# Актуальность исследования

Растущая потребность в автоматизации аналитики и прогнозирования

Необходимость интеграции внешних источников данных

Ограничения традиционных языковых моделей:

- Статичность знаний
- Отсутствие актуальной информации
- Склонность к галлюцинациям

# Цель и задачи исследования

**Цель:**

Разработать систему прогнозирования событий на основе RAG, способную обрабатывать текстовые запросы и генерировать обоснованные прогнозы

**Задачи:**

1. Проанализировать существующие RAG-подходы
2. Реализовать механизм поиска релевантной информации
3. Спроектировать модульную архитектуру системы
4. Интегрировать языковую модель для генерации прогнозов
5. Провести тестирование и оценку эффективности

# Retrieval-Augmented Generation

**RAG = Retrieval + Generation**

**Ретривер** $p_\eta(z|x)$

— поиск релевантных документов

**Генератор** $p_\theta(y_i|x,z,y_{1:i-1})$

— создание ответа на основе найденной информации

**Преимущества RAG**

Динамическое обновление знаний

Фактологическая точность

Прозрачность источников

# Архитектура системы

## Формирование базы знаний

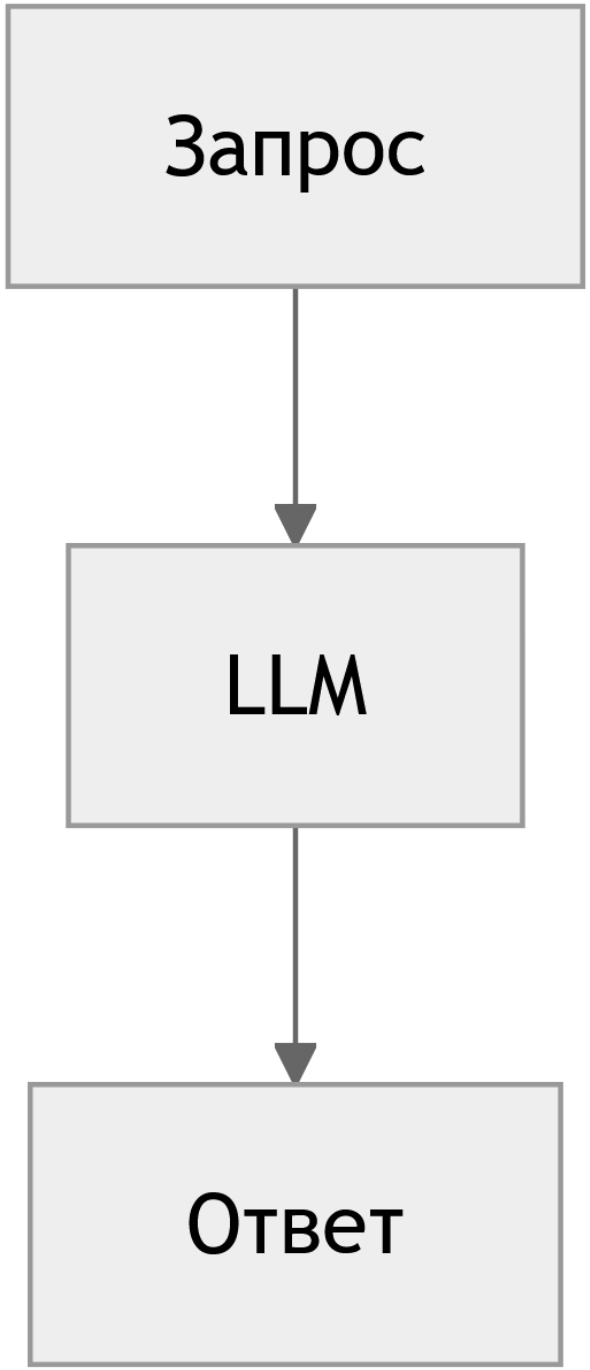
- API The Guardian
- Датасет PROPHET

## Векторизация и индексирование

- Плотная векторизация (SBERT)
- Разреженная векторизация (TF-IDF)

## Поиск и ранжирование

## Генерация ответа

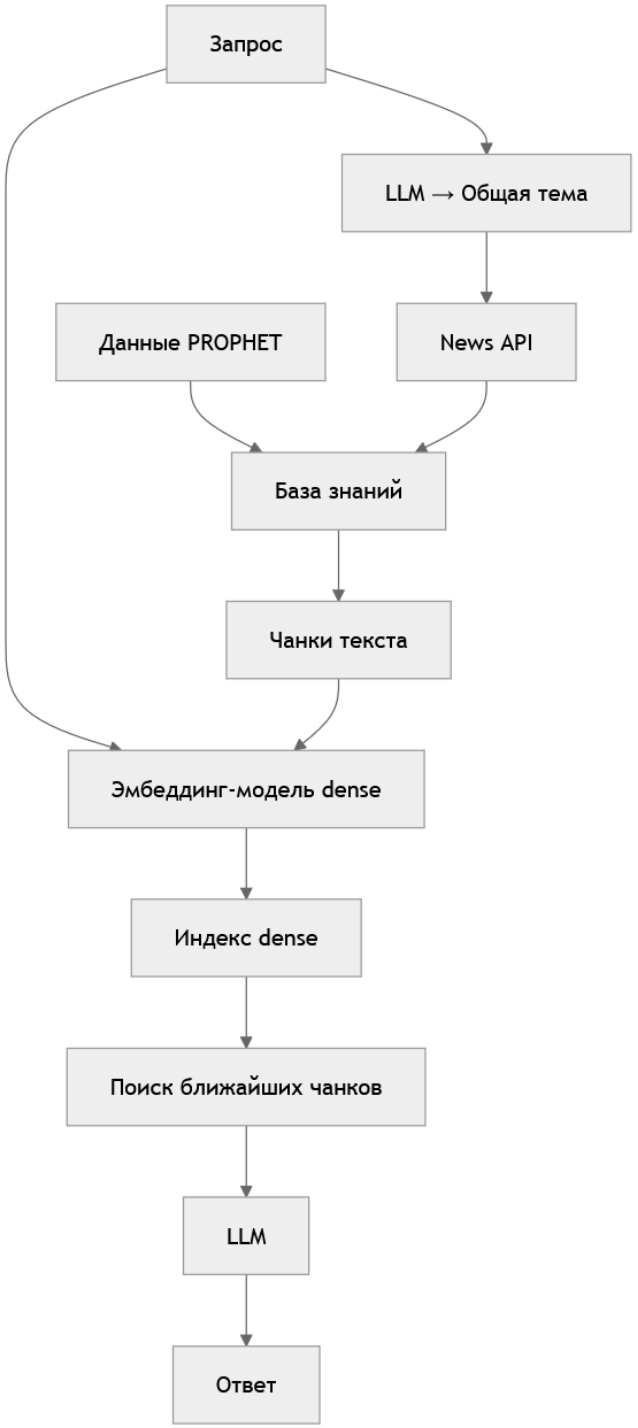



# PlainLLM *Базовый подход*

Языковая модель отвечает только на основе внутренних знаний

Без использования внешних источников

Служит базелайном для сравнения



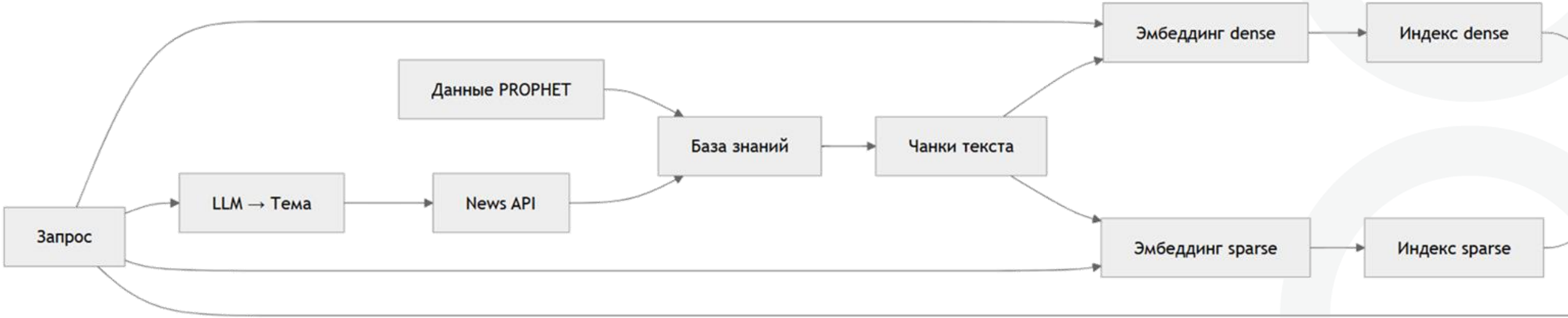

# NaiveRAG *Классический RAG*

Векторизация запроса

Формирование базы знаний

Поиск похожих фрагментов

Генерация прогноза
с контекстом

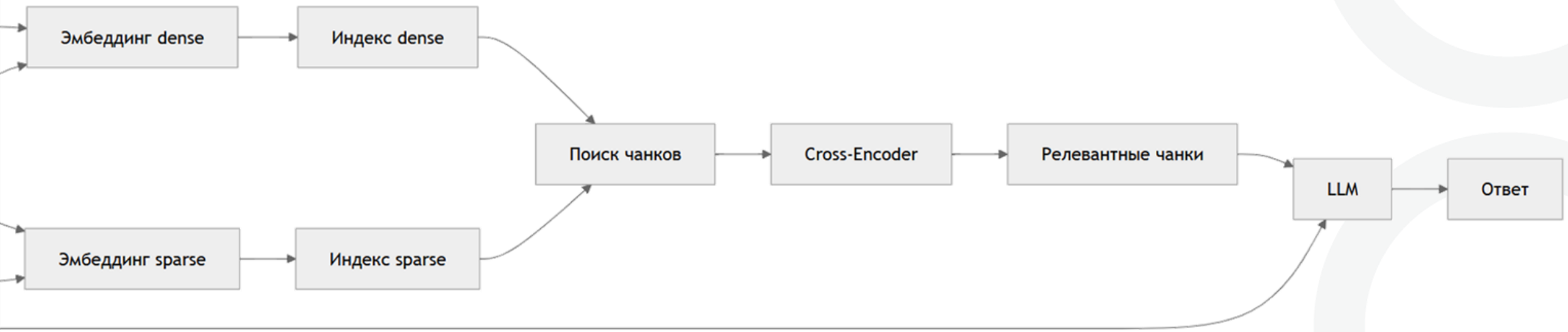



# HybridRAG + Cross-Encoder

Гибридный поиск (плотный + разреженный)

Cross-Encoder переранжирование

Повышенное качество отбора документов




# HybridRAG + Cross-Encoder

Гибридный поиск (плотный + разреженный)

Cross-Encoder переранжирование

Повышенное качество отбора документов

# Реализация системы

**Фреймворки:**

- LangChain
- Groq

**Векторизация:**

- Sentence-Transformers
- FAISS

**Данные:**

- The Guardian API
- PROPHET Dataset

**Модели:**

- Llama-3.1-8b-instant (генерация)
- all-MiniLM-L6-v2 (векторизация)
- ms-marco-MiniLM-L-6-v2 (переранжирование)

# Экспериментальная база

## Источники данных:

**Metaculus**

— научно-ориентированная платформа прогнозирования

**Manifold Markets**

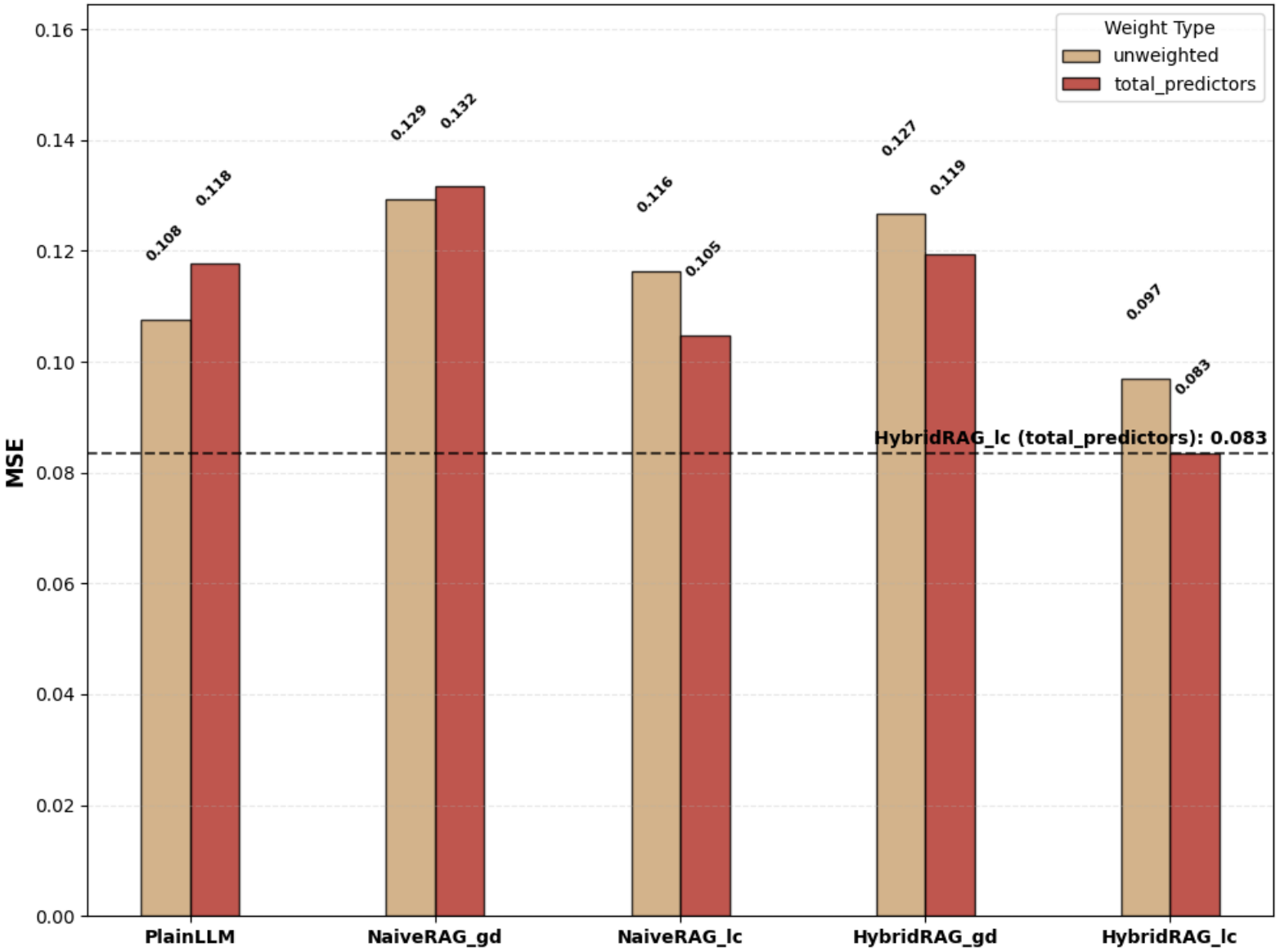
— децентрализованная платформа предсказательных рынков

## Структура датасета:

1. Бинарные вопросы о будущих событиях
2. Фактические исходы событий
3. Коллективные предсказания экспертов
4. Релевантные новостные статьи (отобранные по CIL)

# Результаты

# Предсказание вероятностей (MSE)




## Ключевые результаты:

HybridRAG_lc: **0.097**

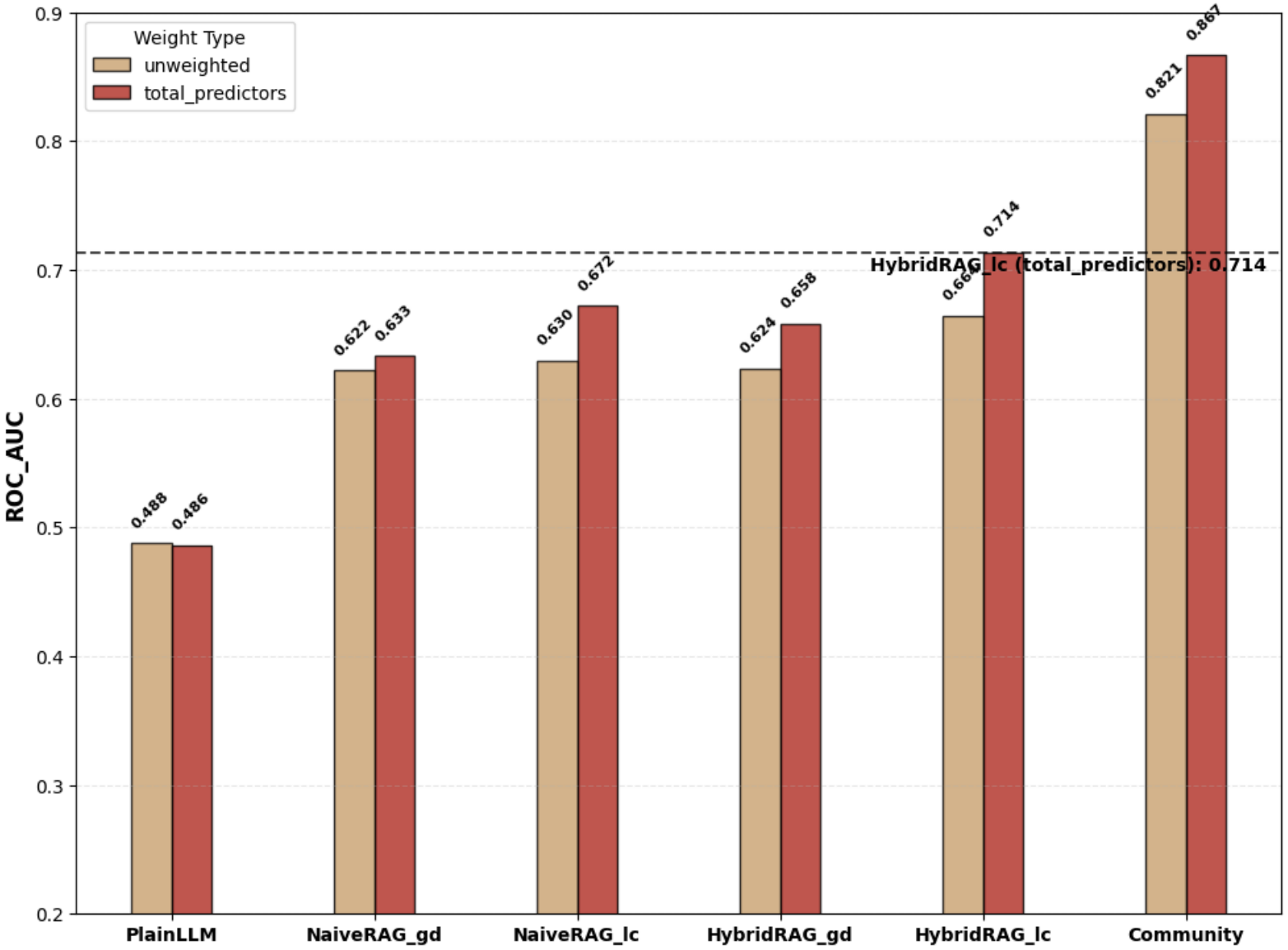
PlainLLM: **0.1075**

NaiveRAG: **0.116-0.129**

Улучшение на **9.8%** относительно базовой LLM

*Comparison of MSE across Models and Weights*

# Бинарная классификация (ROC-AUC)




## Ключевые результаты:

HybridRAG_lc: **0.66-0.71**

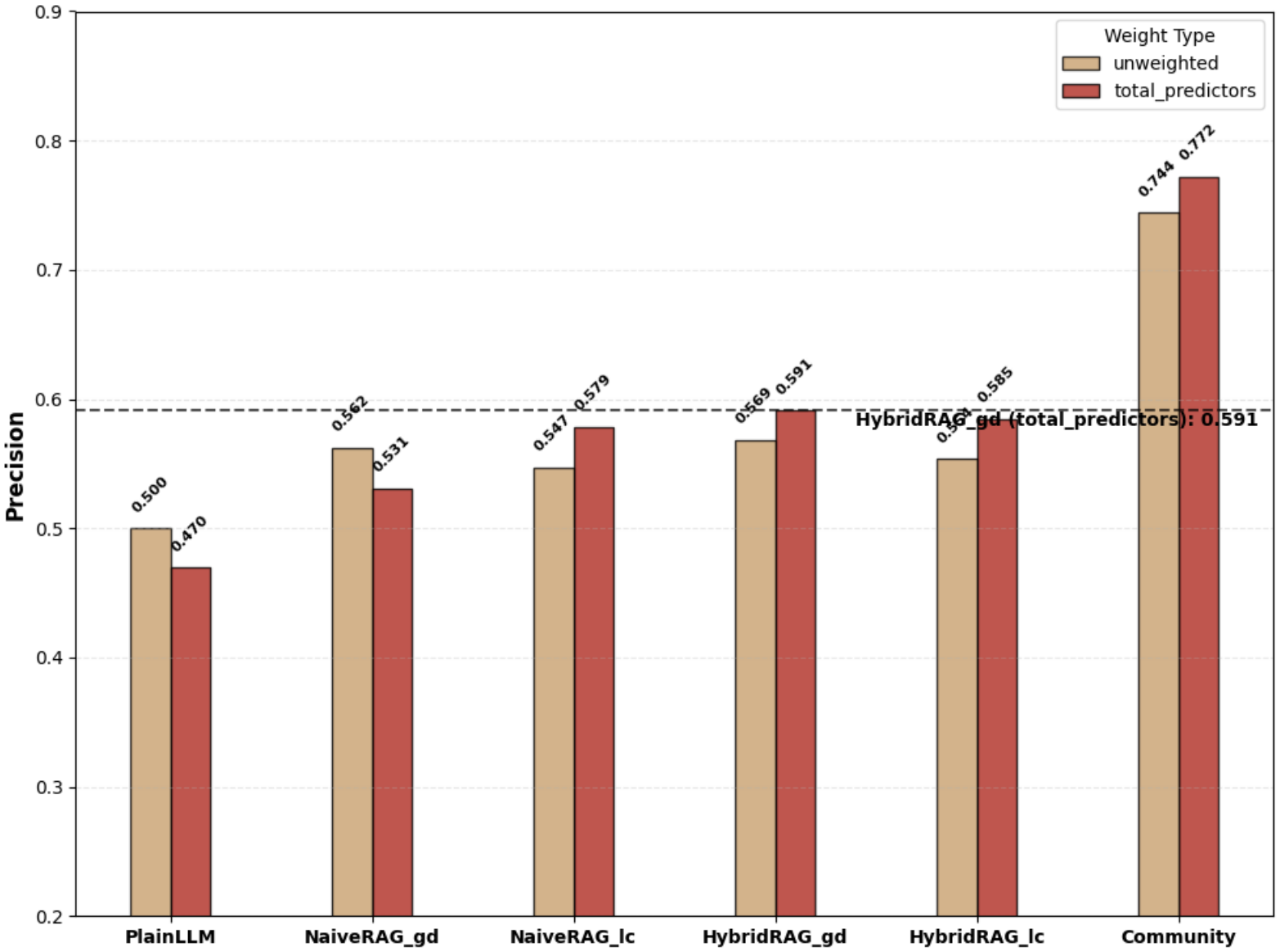
PlainLLM: **0.49**

Community: **0.82-0.87**

Значительное улучшение над базелайном

*Comparison of ROC_AUC across Models and Weights*

# Бинарная классификация (Precision)



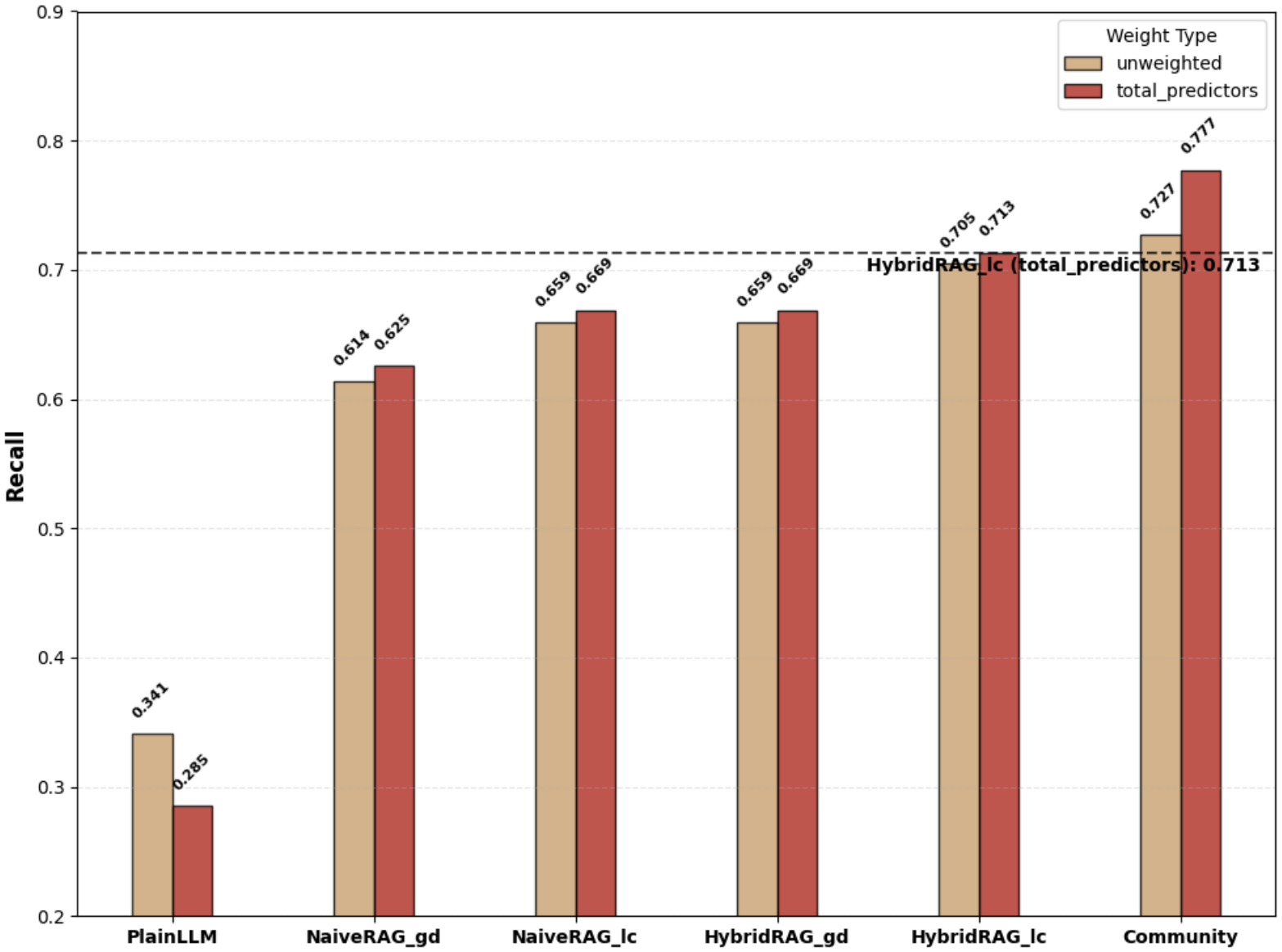

## Ключевые результаты:

Умеренный Precision **0.55-0.58**

Склонность к переоценке рисков

*Comparison of Precision across Models and Weights*

# Бинарная классификация (Recall)

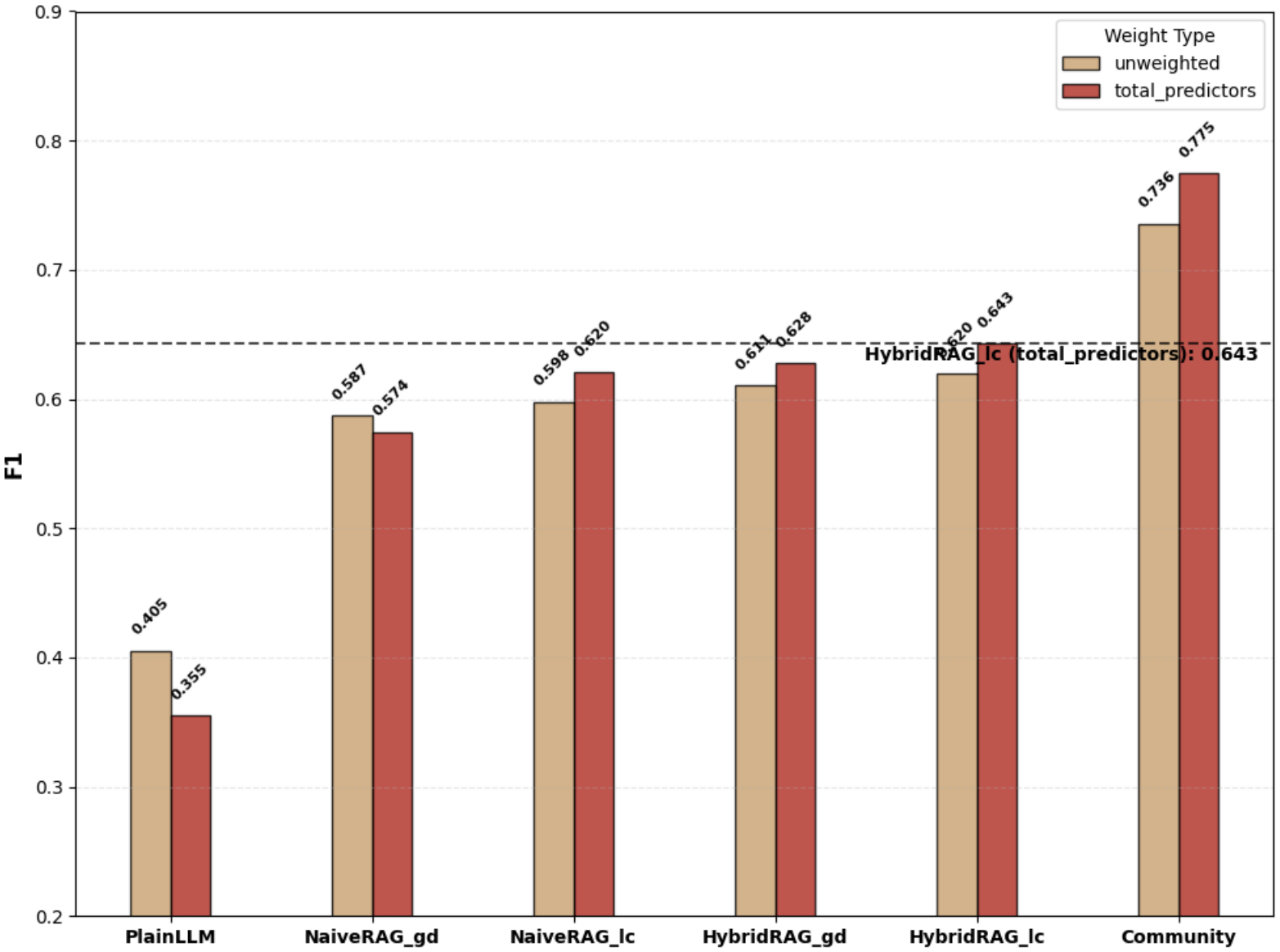



## Ключевые результаты:

Высокий Recall **0.70-0.71**
≈ уровень экспертов

*Comparison of Recall across Models and Weights*

# Бинарная классификация (F1)




*Comparison of of F1 across Models and Weights*

# Анализ

# Ключевые выводы эксперимента

## Основные закономерности:

RAG превосходит базовую LLM — улучшение всех метрик

Локальный корпус эффективнее

Гибридная архитектура лучшая — комбинирование стратегий поиска

## Паттерн производительности:

Высокий Recall ≈ экспертов

Низкий Precision → переоценка событий

Разрыв с коллективными прогнозами

# Ограничения и перспективы

**Выявленные ограничения:**

Переоценка рисков (ложноположительные прогнозы)

Зависимость от качества источников

Отставание от экспертного сообщества

**Направления развития:**

Калибровка вероятностных оценок

Мультиагентные подходы

Улучшение синтеза информации

Адаптивное обновление базы знаний

# Заключение

## Достижения:

Разработана модульная RAG-система

Улучшение MSE на **9.8%,**
ROC-AUC **с 0.49 до 0.66**

Подтверждена эффективность курации данных

## Практическая значимость:

Инструмент поддержки принятия решений

Быстрая обработка текстовой информации

Адаптируемость к различным областям

**СПАСИБО ЗА ВНИМАНИЕ!**

Вопросы?