Министертсво Образования и Науки Республики Казахстан
Казахской Национальный Унивеситет им. Аль-Фараби
Факультет: “Информационные технологии”
Специальность: “Информационные системы”

Тема: “Разработка алгоритма составления синонимов и пополнения записей в БД ”

Группа : ИС-21-4

Студент: Шайдахмет Д. М.

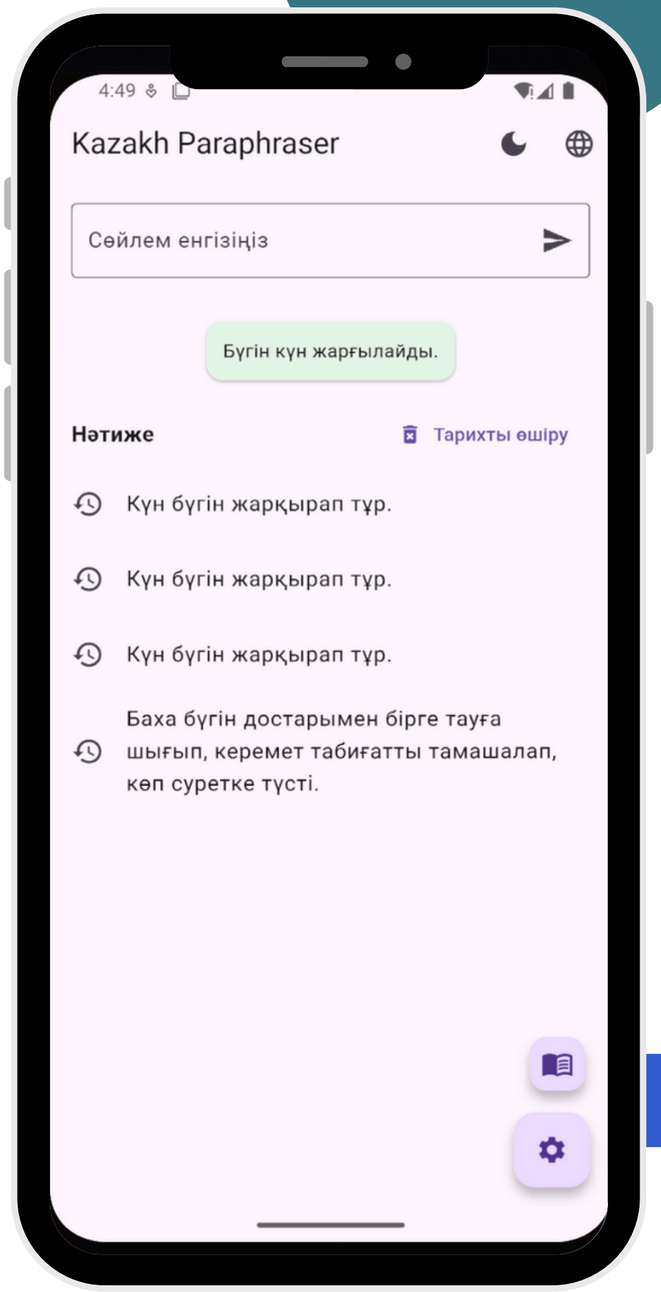
Руководитель: Шормакова А. Н. PhD., и.о. доцента

Алматы 2025

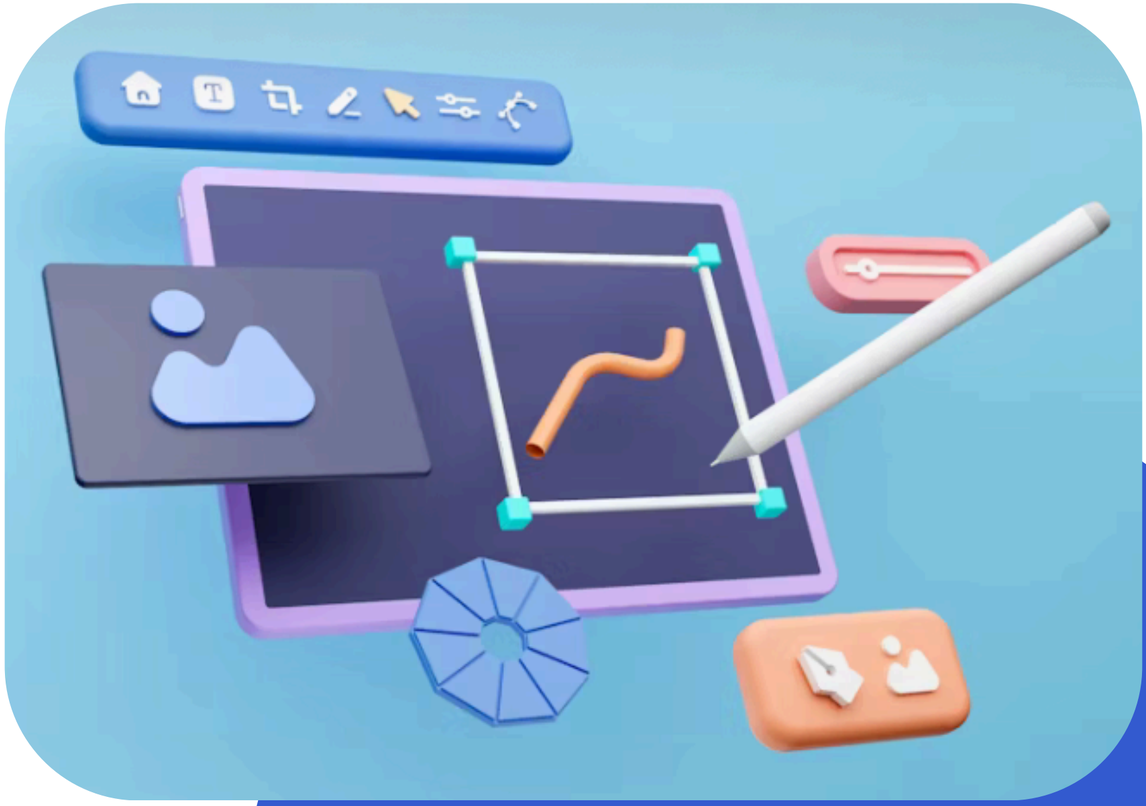
СОДЕРЖАНИЕ:

ВСТУПЛЕНИЕ

На сегодняшний день казахский язык слабо поддерживается в современных системах обработки текста. Существующие решения не справляются с точным перефразированием и не учитывают особенности языка.
В рамках проекта было создано удобное приложение, которое позволяет пользователю быстро вводить казахский текст, получать синонимизированный вариант и сохранять результат в базу данных.
Система обеспечивает быстрый доступ, простоту использования и может применяться в образовании, переводе и других сферах, где важна работа с текстами на казахском языке.



АКТУАЛЬНОСТЬ, НОВИЗНА И ЗНАЧИМОСТЬ



01

АКТУАЛЬНОСТЬ

- Казахский язык слабо представлен в современных NLP-системах.
- Большинство существующих решений не учитывают его грамматическую и синтаксическую структуру.
- Это ограничивает применение автоматизированной обработки текста в образовании, науке и повседневной практике.

02

НОВИЗНА

- Впервые реализован комплексный подход, сочетающий генерацию синонимов с сохранением результатов в базу данных.
- Разработана система, ориентированная именно на казахский язык и его особенности.
- Использован авторский корпус из 134 000 предложений.

03

ЗНАЧИМОСТЬ

- Система способствует цифровизации казахского языка.
- Может применяться в образовании, переводе, разработке лингвистических ресурсов.
- Создаёт базу для дальнейших исследований в области казахской компьютерной лингвистики.

АНАЛИЗ АНАЛОГОВ

СРАВНЕНИЕ С:

- ChatGPT
- Quillbot
- Paraphraser.io



Недостатки:

- Нет поддержки казахского языка
- Нет доступа к базе данных
- Не адаптированы под локальные задачи

```python
import pandas as pd
from transformers import T5Tokenizer, T5ForConditionalGeneratic
from datasets import Dataset
```

```python
def preprocess_function(examples):
    inputs = ["paraphrase: " + text for text in examples["src"]]
    targets = examples["trg"]
    model_inputs = tokenizer(inputs, padding="max_length", truncation=True, max_length=128)
    labels = tokenizer(targets, padding="max_length", truncation=True, max_length=128).input_ids
    model_inputs["labels"] = labels
    return model_inputs
```

```python
tokenized_dataset = dataset.map(preprocess_function, batched=True)
```

```python
training_args = TrainingArguments(
    output_dir="./paraphrase_model",
    evaluation_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=100,
    weight_decay=0.01,
    save_total_limit=2,
    save_strategy="epoch"
)
```

```python
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_dataset,
    eval_dataset=tokenized_dataset,
    tokenizer=tokenizer
)
```

МЕТОДОЛОГИЯ

01 Архитектура: T5 (Text-to-Text Transformer)

02 Схема: "Вводное предложение" -> "Перефраз"

03 Fine-tuning модели на задаче перефразирования

03 Использование префикса "paraphrase: "

ДАТАСЕТ

01 134 000 пар предложений

02 Источники: книги, сайты, диалоги

03 Очистка и нормализация

04 Подготовка: токенизация, разметка, разбиение

	src	trg
0	«Single Ladies» музыкалық бейнебаяны бүкіл әле...	The Toronto Star мәлімдеуінше, 'Single Ladies'...
1	«Single Ladies» музыкалық бейнебаяны бүкіл әле...	Single Ladies' музыкалық бейнебаяны бірінші бе...
2	«Single Ladies» музыкалық бейнебаяны бүкіл әле...	Toronto Star мәліметі бойынша, 'Single Ladies'...
3	«Single Ladies» музыкалық бейнебаяны бүкіл әле...	Toronto Star хабарлағандай, 'Single Ladies' му...
4	«Single Ladies» музыкалық бейнебаяны бүкіл әле...	Sen The Toronto Star, 'Single Ladies' музыкалы...
...	...	...
131239	Ең ақылды жануарлардың арасында кейбір құстар,...	Құстардың кейбір түрлері, әсіресе корвидтер ме...
131240	Ең ақылды жануарлардың арасында кейбір құстар,...	Ең ақылды жануарлардың арасында кейбір құстар,...
131241	Кейбір құстар, әсіресе корвидтер мен тотықуста...	Құстардың кейбір түрлері, әсіресе корвидтер ме...
131242	Кейбір құстар, әсіресе корвидтер мен тотықуста...	Ең ақылды жануарлардың арасында кейбір құстар,...
131243	Құстардың кейбір түрлері, әсіресе корвидтер ме...	Ең ақылды жануарлардың арасында кейбір құстар,...

131244 rows × 2 columns

```
tokenized_dataset = dataset.map(preprocess_function, batched=True)
```

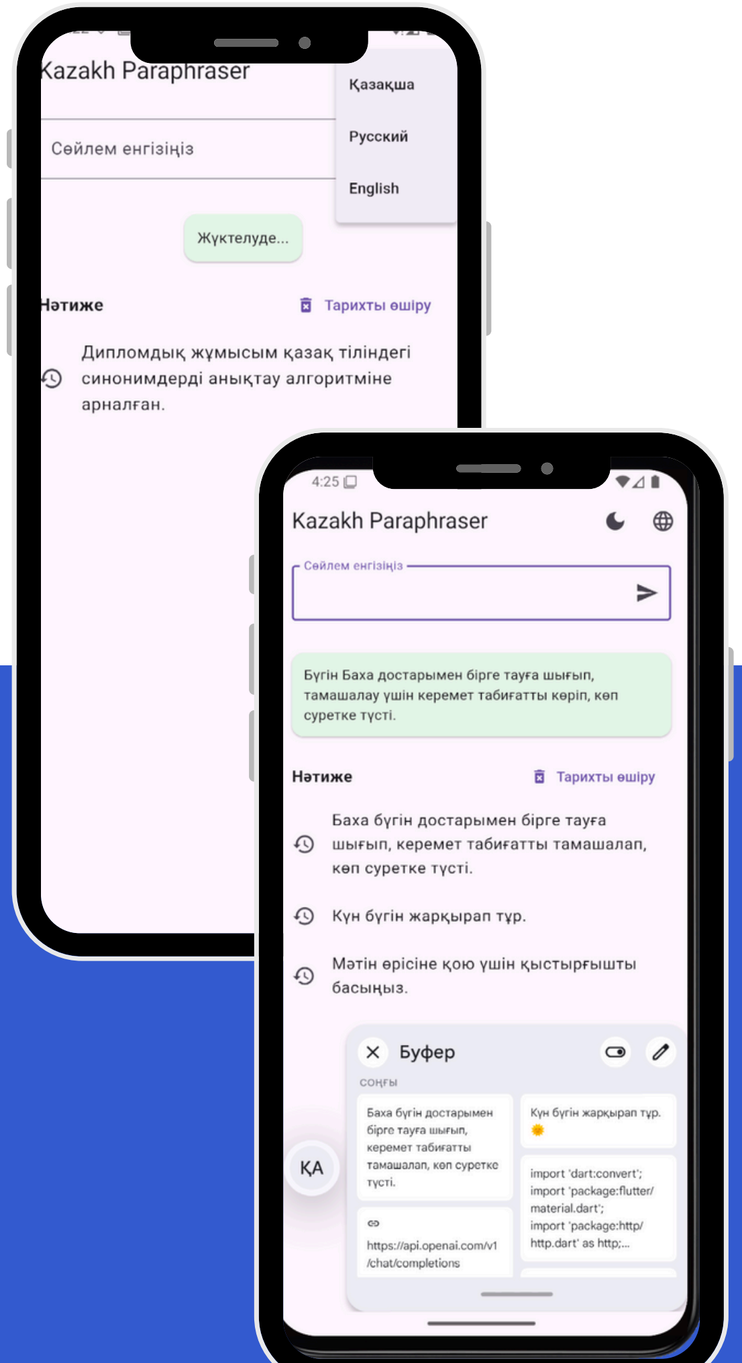
СЕРВЕРНАЯ ЧАСТЬ

Реализация на Python + FastAPI

Функции:

- Прием запросов
- Генерация синонимов через T5
- Возврат результата
- Запись в БД

КЛИЕНТСКАЯ ЧАСТЬ



- **Приложение на Flutter**
- **UI: текстовое поле, кнопка, результат**
- **Возможность отправки результата в БД**
- **История и сохранение данных**

РЕЗУЛЬТАТЫ И ПРИЕМУЩЕСТВА

Результаты:

- Среднее время отклика системы составляет около 1 секунды.
- Модель генерирует семантически точные, грамматически корректные синонимы.
- Создана связанная инфраструктура: сервер — модель — база данных — клиентское приложение.
- Пример:
  - Оригинал: Мен мектепке бардым
  - Синоним: Мен оқу орнына аттандым

Приемущества

- Среднее время отклика системы соПоддержка казахского языка, чего нет в большинстве аналогов.
- Простота и удобство интерфейса для конечного пользователя.
- Возможность постоянного пополнения базы данных синонимами.
- Адаптируемость решения под другие задачи: упрощение текста, перевод, генерация вопросов.
- Высокая точность благодаря использованию качественного авторского корпуса.

СПАСИБО ЗА ВНИМАНИЕ