



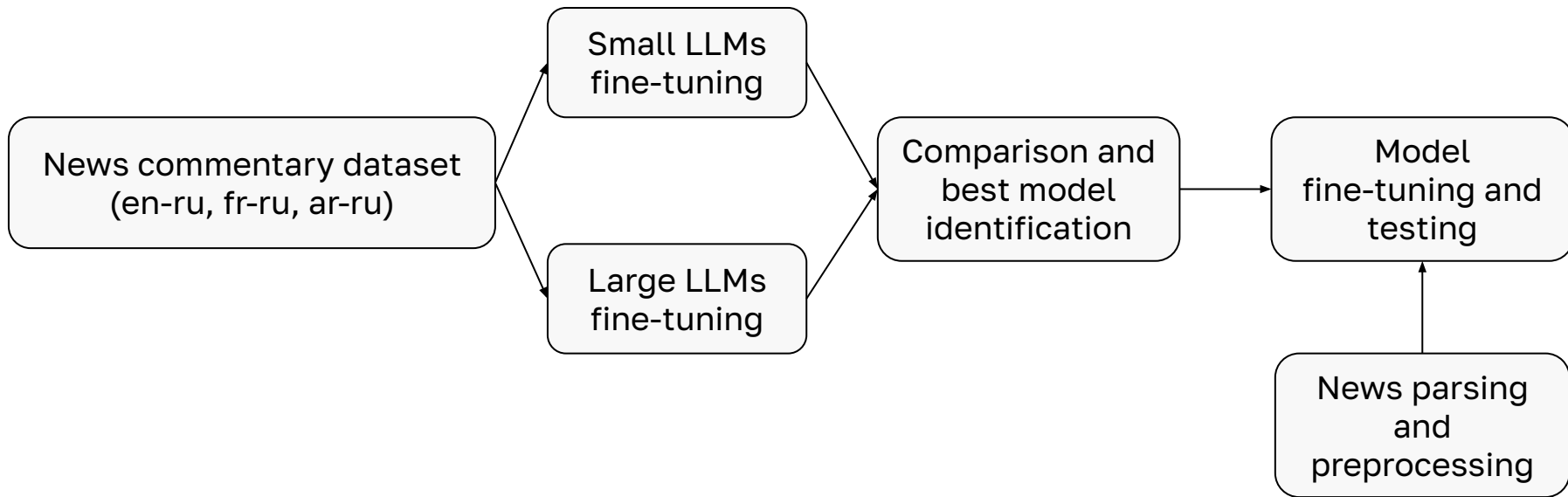
**IITMO**

# **Multi Language Translation of News into Russian**

Students:  
Lepekhina Anfisa,  
Iufriakova Anastasiia,  
J4234c

- translation into russian is still challenging
- domain-specific problem
- real-world relevance
- access to information

# Work pipeline



# Data overview

**OPUS News Commentary data** – main dataset for models training

- OPUS is a growing collection of translated texts from the web

Languages	Dataset size
arabic-russian	84.5k
french-russian	161k
english-russian	190k

*English example:*

"en": Gold prices are extremely sensitive to global interest-rate movements.

"ru": "Цены на золото чрезвычайно чувствительны к мировым движениям процентных ставок."

# Models reviewed

## LLMs for working in Kaggle:

- google/mt5-base - **580M** parameters
- facebook/m2m100\_418M - **418M** parameters

## LLMs for final results on TeslaA100:

- google/mt5-large - **1.2B** parameters
- facebook/m2m100\_1.2B - **1.2B** parameters

## MT5

- a multilingual variant of T5 (Text-to-Text Transfer Transformer)
- was pre-trained on a new Common Crawl-based dataset covering 101 languages

## M2M100

- a multilingual encoder-decoder (seq-to-seq) model
- non-English-Centric model
- was pre-trained on large-scale Many-to-Many dataset for 100 languages (mining + back translation strategy)
- evaluated on publicly available datasets (WMT, WAT, IWSLT, etc.)

# Chosen metrics: SacreBLEU and ROUGE

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-\frac{r}{c}} & \text{if } c \leq r \end{cases}, \quad BLEU = BP \cdot (\sum_{n=1}^N w_n \log p_n)$$

BP - brevity penalty to control candidate length

N - n-gram order (usually =4)

w\_n - weights

p\_n - precision for n-grams in the candidate compared to the reference

$$ROUGE\_N = \frac{\sum_{S \in \{references\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{references\}} \sum_{gram_n \in S} Count(gram_n)}$$

# Results: M2M100\_1.2B

- Metrics were evaluated on 1000 samples

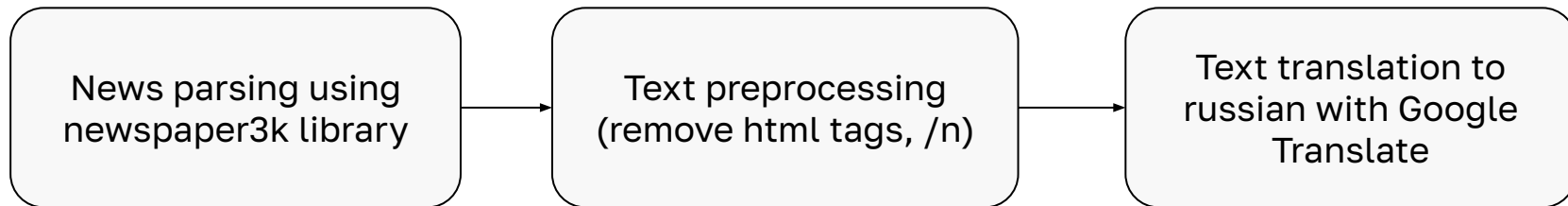
Language	BLEU (previous)	BLEU	ROUGE_1	ROUGE_2
English	28.18	<b>29.39</b>	0.164	<b>0.087</b>
French	18.67	19.38	0.130	0.048
Arabic	13.67	16.27	<b>0.166</b>	0.067



# Testing models on real data

Sources for news parsing for final model testing:

- <https://www.skynewsarabia.com/>
- <https://edition.cnn.com/>
- <https://www.francetvinfo.fr/>



# Results: test dataset

- Metrics were evaluated on 50 samples for each language

Language	BLEU	ROUGE_1	ROUGE_2
English	27.68	0.507	0.166
French	17.00	0.473	0.202
Arabic	20.41	0.351	0.134

- **in most cases** the translation of google translator and model are **the same**
- in cases when we can see difference between translation, google translator provide **more detailed translation**

# Examples of translation

## Model translation

EN

Апелляционный суд восстановил приказ GAG, запрещающий бывшему президенту Дональду Трампу делать публичные заявления **о сотрудниках зала суда** в ходе продолжающегося судебного разбирательства в 250 миллионов долларов США.

AR

Президент Объединенных Арабских Эмиратов **объявил создание** частного фонда, посвященного «решениям» **перед лицом изменения климата**, открыв саммит мировых лидеров **28 -я конференция** Организации Объединенных Наций по климату.

FR

Во Франции **насчитывается** около 200 000 ВИЧ-позитивных людей, которые живут благодаря эффективным методам лечения ВИЧ, но иногда с довольно **серьезными** побочными эффектами.








## Google translation



Апелляционный суд вновь установил ордер, запрещающий бывшему президенту Дональду Трампу публично заявлять **о представителях палаты суда** во время продолжающегося судебного процесса по мошенничеству в размере 250 миллионов долларов США.

Президент Объединенных Арабских Эмиратов **объявил о создании** частного фонда, посвященного «решениям» **по борьбе с изменением климата**, открыв саммит мировых лидеров **на 28-й Конференции** Организации Объединенных Наций по изменению климата.

Во Франции **существует** около 200 000 ВИЧ -позитивных людей, которые живут благодаря эффективным методам лечения ВИЧ, но иногда с довольно **тяжелыми** побочными эффектами.


- Facebook **M2M100** and Google **mT5** LLMs were **fine-tuned** for multi-language translation task
- Chosen **models were compared** according to selected quality **metrics** experimentally
- News **parser** was **implemented** to test models on real data
- The **best model was tested** on English, French and Arabic parsed news
- The **results** of the model are **close to real** translations, and in some cases are even **better**

 <b>anafisa</b> add: model fine-tunning	4d6d807 · 18 minutes ago	 6 Commits
 parser	add: news parser	24 minutes ago
 test	add: test on real data	23 minutes ago
 train	add: model fine-tunning	18 minutes ago
 LICENSE	Initial commit	1 hour ago
 README.md	Update README.md	1 hour ago

 README  MIT license

## Large Multi-Language Models for News Translation

- In this repo you may find examples **how to fine-tune Large Language Models (LLM)** and apply them to the real task of **news translation**.
- Also in this repo we provide **news parser**, so you can easily parse any news web page you want (for example CNN, BBC news) and test how pre-trained LLM would **translate parsed real news**.



<https://github.com/anafisa/Multi-Language-Translation-of-News>



**I/TMO**

**THANK YOU  
FOR YOUR TIME!**