TASK:

The task involves utilizing multilingual LLMs to conduct few-shot inference for NLI and Relation Extraction tasks, with examples chosen from one language and query input from another, avoiding exact translations and selection strategies for in-context examples, including random, task-aligned, semantic-aligned and semantic&task-aligned settings.

Approach:
We utilized the T5 model as our multilingual LLM. In the random-shot approach, we randomly selected k instances from the validation set to obtain the top k few-shot instances. Semantic alignment involved computing embeddings of test instances and comparing them with validation set instances to select the top k instances. Task alignment included incorporating a task aligner sentence, such as: "In French, vrai means true, inconnu means false, and faux means unknown." For semantic&task-aligned strategies, the task aligner was added after the few-shot prompt.

Given that the SMILER dataset has labels in English for all languages, ensuring uniformity across test instances, it does not necessitate a task aligner in the prompt.

Results:
For XNLI dataset:

Random few shot
Precison for {few shot language}-{test instance language}

```
Precision for en-fr: 0.10247208931419456
Precision for en-ru: 0.09584470730967545
Precision for fr-en: 0.17318007662835247
Precision for fr-ru: 0.08811188811188812
Precision for ru-en: 0.06666666666666667
Precision for ru-fr: 0.0782122905027933

Overall precision across all language pairs: 0.10074795308892844
```

Semantic few-shot:
Precison for {few shot language}-{test instance language}

```
Precision for ru-fr: 0.11166666666666668
Precision for en-fr: 0.22447828539199097
Precision for en-ru: 0.11166666666666668
Precision for fr-en: 0.3003684312584836
Precision for fr-ru: 0.19557823129251703
Precision for ru-en: 0.44612794612794615
Precision for ru-fr: 0.11166666666666668

Overall precision across all language pairs: 0.2316477045673785
```

Task align few shot:

Precison for {few shot language}-{test instance language}
Precision for ru-fr: 0.11166666666666668
Precision for en-fr: 0.056131260794473226
Precision for en-ru: 0.05946275946275946
Precision for fr-en: 0.26943005181347146
Precision for fr-ru: 0.26844919786096255
Precision for ru-en: 0.08417085427135679
Precision for ru-fr: 0.11166666666666668

Overall precision across all language pairs: 0.14155179847828173

Semantic-taks align few shot:
Precison for {few shot language}-{test instance language}
Precision for en-fr: 0.24444444444444446
Precision for en-ru: 0.17695140863272163
Precision for fr-en: 0.26760031579308685
Precision for fr-ru: 0.20607136817182478
Precision for ru-en: 0.34652014652014657
Precision for ru-fr: 0.27777777777777773

Overall precision across all language pairs: 0.25322757689000036

For SMILER dataset:

Random few shot:
Precison for {few shot language}-{test instance language}
Precision for en-fr: 0
Precision for en-ru: 0
Precision for fr-en: 0.01
Precision for fr-ru: 0
Precision for ru-en: 0.03
Precision for ru-fr: 0

Semantic few shot:
Precison for {few shot language}-{test instance language}
Precision for en-fr: 0.02
Precision for en-ru: 0
Precision for fr-en: 0
Precision for fr-ru: 0
Precision for ru-en: 0.03
Precision for ru-fr: 0

Analysis:

For the XNLI dataset, implementing a task-aligned approach enhances precision by ensuring consistent label information across languages, resulting in a notable 4% overall improvement. However, the effectiveness of semantic alignment versus semantic-task alignment presents a mixed picture, exhibiting varied precision across different language pairs. While certain pairs, like en-fr and fr-ru, benefit from semantic alignment, others show minimal or negative effects. Notably, the task-aligned strategy aids in improving results even within semantic few-shot settings.

Due to LLMs being predominantly trained on English, the performance for English test inputs surpasses that of French and Russian. Additionally, given the closer linguistic association between French and English, few-shot learning in French contributes more to enhancing English performance compared to Russian. The augmentation of model understanding through label information significantly enhances overall performance.

In the SMILER dataset, where labels for Russian and French are also in English, uniformity across languages precludes the feasibility of task alignment. However, the exceptionally low results for the SMILER dataset can be attributed to its high label count of 32, surpassing typical NLP classification datasets like XNLI. Consequently, smaller models like T5 and XGLM struggle to provide accurate results due to the increased contextual demands.