# PARROT : Translating During Chat Using Large Language Models

In this work,

- Parrot framework was proposed to enhance and regulate the translation abilities during chat based on open-sourced LLMs(LLaMA -7b, BLOOMZ-7b-mt) and human written translation and evaluation data.
- Parrot reformulates translation data into the instruction-following style, and introduces a “Hint” field for incorporating extra requirements to regulate the translation process.
- Accordingly, three instruction types for finetuning Parrot models were proposed, which are translation instruction, contrastive instruction and error-guided instruction.

The three distinct instruction types proposed are :

- **Translation Instruction:** Asks the  LLMs to generate translations based on source sentences. Stanford Alpaca was followed to transform bilingual sentence pairs into the instruction-following format.
- **Contrastive Instruction:** Asks the LLMs to generate the translations based on the preference. Multiple different translations for each source were acquired by the systems submitted to WMT competitions.
- **Error-Guided Instruction:** Asks the LLMs to generate the translations with human annotated errors as the hint.

The first instruction guarantees the basic translation ability of LLMs while the latter two regulate the LLMs to align with human feedbacks.

Table 1: Instruction types for finetuning PARROT models.

| Type | Instruction |
|---|---|
| | Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.\n\n |
| Translation | ### Instruction:<br>Translate the following sentences from Chinese to English.\n\n<br>### Input:<br>检查情况显示，市场销售的粮油、肉类、水果、蔬菜、蛋奶等生活必需品供应充足，商品价格基本稳定，未发现严重违法违规行为，市场经营秩序总体平稳。\n\n<br>### Response:The inspection results showed that there was an adequate supply of daily necessities, including grain, oil, meat, fruit, vegetable, milk, and eggs in the market and commodity prices basically remain stable, the administration found no serious offensive and noncompliant conducts, and the market order remains stable on the whole. |
| Contrastive | ### Instruction:<br>Translate the following sentences from Chinese to English.\n\n<br>### Input:<br>检查情况显示，市场销售的粮油、肉类、水果、蔬菜、蛋奶等生活必需品供应充足，商品价格基本稳定，未发现严重违法违规行为，市场经营秩序总体平稳。\n\n<br>### Hint: We prefer to translate it to\n\n<br>### Response: The inspection results showed that there was an adequate supply of daily necessities, including grain, oil, meat, fruit, vegetable, milk, and eggs in the market and commodity prices basically remain stable, the administration found no serious offensive and noncompliant conducts, and the market order remains stable on the whole. rather than The results of the inspection indicate the sufficient supply of living necessities on marketing including cereals and oils, meat, fruits, vegetables, eggs and milk, and the basically stabilized commodity price. The inspection hasn't found serious violation of laws and regulations. The market order is stable on an overall basis. |
| Error-Guided | ### Instruction:<br>We are translating the following sentences from Chinese to English. \n\n<br>### Input:<br>检查情况显示，市场销售的粮油、肉类、水果、蔬菜、蛋奶等生活必需品供应充足，商品价格基本稳定，未发现严重违法违规行为，市场经营秩序总体平稳。\n\n<br>### Hint: A translation with major accuracy/mistranslation errors could be\n\n<br>### Response:The results of the inspection indicate the sufficient supply of living necessities <v>on marketing</v> including cereals and oils, meat, fruits, vegetables, eggs and milk, and the basically stabilized commodity price. The inspection hasn't found serious violation of laws and regulations. The market order is stable on an overall basis.<br>### Hint: A translation with minor fluency/grammar errors could be\n\n<br>### Response:The results of the <v>inspection</v> indicate the sufficient supply of living necessities on marketing including cereals and oils, meat, fruits, vegetables, eggs and milk, and the basically stabilized commodity price. The inspection hasn't found serious violation of laws and regulations. The market order is stable on an overall basis. |

## TRAINING DATA

- **Alpaca Data**
- **WMT Validation Data :** The newstest2017- 2020 of Chinese⇔English (i.e., Zh⇔En) and German⇔English (i.e., De⇔En) tasks, which consist of 51.2K sentence pairs for all the four directions.

- **MQM Human  Evaluation Data:** Here,human evaluation data comes from the Multidimensional Quality Metrics (MQM) datasets, which annotate the different translation errors (e.g., major accuracy/mistranslation, minorfluency/grammar) of top WMT systems.

**TEST DATA**

- **Flores Subset**
- **WMT22 Test Sets**

**MODEL TRAINING**

LLaMA-7b and BLOOMZ-7b-mt were adopted as base models and were finetuned to following variants:

- **Alpaca :** Finetuned only on Alpaca multi-task data set.
- **Parrot :** Finetuned on both the Alpaca multitask dataset and WMT validation data.
- **Parrot-Hint :** Finetuned on the Alpaca multitask dataset, WMT validation data, MQM human evaluation data as well as the automatically assessed data.
- **LLMs - LoRA :**  The above LLMs finetuned by low-rank adaptation (LoRA), which contains only around 4.2M tunable parameters with the default hyper-parameters.

**RESULTS**

Table 2: Ablation study of key factors on Flores En⇒De subset with Alpaca based on LLaMA-7b.

| Prompt | Instruct. | Search | BLEU | COMET |
|---|---|---|---|---|
| no-input | TP1 | sample | 20.09 | 80.03 |
| | | beam 4 | 22.19 | 79.13 |
| | TP3 | sample | 19.43 | 79.00 |
| | | beam 4 | 21.52 | 79.08 |
| input | TP1 | sample | 21.00 | 79.51 |
| | | beam 4 | **23.32** | **80.56** |
| | TP3 | sample | 19.33 | 78.68 |
| | | beam 4 | 20.64 | 80.07 |

(1) The prompt-input performs slightly better than prompt-no-input though the gap is marginal.
(2) The TP1 instruction works better on Alpaca than TP3 which is different from that on ChatGPT.
(3) Generally, beam search outperforms sampling significantly, especially in terms of BLEU score.
Therefore, they use prompt-input + TP1 + beam search as the default setting for inference.

Table 3: Translation performance of LLMs on Flores subsets and WMT22 test sets.

| System | De⇒En | | En⇒De | | Zh⇒En | | En⇒Zh | |
|---|---|---|---|---|---|---|---|---|
| | BLEU | COMET | BLEU | COMET | BLEU | COMET | BLEU | COMET |
| **Flores Subsets** | | | | | | | | |
| Google | 45.04 | 88.79 | 41.16 | 88.61 | **31.66** | **87.71** | 43.58 | **88.42** |
| DeepL | **49.23** | **89.70** | 41.46 | 89.03 | 31.22 | 87.39 | **44.31** | 88.11 |
| ChatGPT | 43.71 | 89.10 | 38.87 | 88.14 | 24.73 | 85.81 | 38.27 | 86.99 |
| GPT-4 | 46.00 | 89.31 | **45.73** | **89.28** | 28.50 | 87.42 | 42.50 | 88.40 |
| *Base Model: LLaMA-7b* | | | | | | | | |
| Vanilla | 3.43 | 60.15 | 2.41 | 49.03 | 1.86 | 53.73 | 0.05 | 47.64 |
| Alpaca | 36.66 | 86.87 | 23.32 | 80.56 | 15.13 | 81.20 | 9.81 | 58.69 |
| Alpaca-LoRA | 40.77 | 87.70 | 24.64 | 84.07 | 16.42 | 81.59 | 14.51 | 70.52 |
| PARROT | 41.34 | 87.75 | 28.58 | 83.30 | 19.56 | 83.17 | 24.77 | 79.90 |
| PARROT-HINT | 41.07 | 87.97 | 30.83 | 84.38 | 19.22 | **83.96** | 25.84 | 80.15 |
| + Infer w/ Prefer. | 38.10 | 87.60 | 23.07 | 83.97 | 18.69 | 83.10 | 22.54 | 80.12 |
| + Infer w/ No Err. | 42.25 | **88.73** | **32.10** | 84.94 | **21.57** | 83.72 | **27.44** | **81.88** |
| PARROT-HINT-LoRA | **43.82** | 88.36 | 29.02 | 84.90 | 16.94 | 80.66 | 14.83 | 71.53 |
| + Infer w/ No Err. | 42.01 | 88.08 | 29.89 | **85.40** | 17.43 | 81.38 | 19.89 | 76.78 |
| **WMT22 Test Sets** | | | | | | | | |
| WMT22 Winners | 33.70 | -/- | 38.40 | -/- | 33.50 | -/- | 54.30 | -/- |
| *Base Model: LLaMA-7b* | | | | | | | | |
| Vanilla | 2.90 | 52.81 | 1.63 | 45.36 | 1.25 | 50.34 | 0.38 | 46.38 |
| Alpaca | 27.82 | 82.33 | 20.17 | 78.13 | 14.24 | 74.00 | 10.44 | 62.10 |
| Alpaca-LoRA | 28.98 | **83.23** | 22.19 | 81.33 | 16.13 | 75.63 | 16.34 | 70.63 |
| PARROT | 26.61 | 82.57 | 24.03 | 80.43 | 18.10 | 75.38 | 27.08 | 78.45 |
| PARROT-HINT | 27.33 | 82.47 | 24.68 | 81.22 | 18.90 | 75.26 | 28.15 | 79.31 |
| + Infer w/ No Err. | 27.38 | 82.47 | **26.14** | 81.67 | **20.23** | **75.90** | **30.33** | **80.34** |
| PARROT-HINT-LoRA | 28.85 | 82.88 | 24.04 | 81.47 | 18.27 | 74.77 | 19.93 | 73.70 |
| + Infer w/ No Err. | **29.86** | 83.08 | 24.89 | **81.68** | 19.20 | 75.00 | 20.76 | 74.51 |
| *Base Model: BLOOMZ-7b-mt* | | | | | | | | |
| Vanilla | 11.15 | 66.38 | 2.24 | 47.07 | 0.25 | 57.62 | 0.94 | 47.67 |
| Alpaca | 17.60 | 73.02 | 3.11 | 44.51 | 13.04 | 76.40 | 23.96 | 81.88 |
| PARROT-HINT | 23.15 | 77.62 | 20.01 | 72.73 | 21.46 | 78.57 | 32.40 | **83.69** |
| + Infer w/ No Err. | **24.96** | **78.09** | **20.56** | **73.62** | **22.72** | **79.00** | **34.58** | 83.54 |

**OBSERVATIONS:**

- **Instruction tuning exploits the potential of vanilla LLMs for machine translation:**
    - vanilla LLaMA-7b performs badly on theFlores subsets, the model tends to generate very long sentences which makes the generated text not faithful to the source sentences and also not grammatically correct.

    - Tuning LLaMA-7b on the Alpaca multi-task dataset (i.e., Alpaca) performs much better on translation.
    - However, the best performance is mainly observed on high-resource directions like De⇒En, due to the dominant language of Alpaca dataset in English.
    - Further introducing a small amount of translation instructions (i.e., PARROT) in the four language directions can significantly improve the performance, especially for En⇒Zh, in which Chinese was unseen in the pretraining of LLaMA models.
- **Learning from low-quality translations annotated by human is also important:**
    - With no hint for inference, PARROT-HINT outperforms PARROT slightly on translation directions from English to other langauges (i.e., En⇒De, En⇒Zh).
    - However, when asking PARROT-HINT to generate translations with no error, the performance can be significantly improved across translation directions and test sets.
    - PARROT-HINT learns the relationship between errors and translations by error-guided instruction, such that it can avoid the translation errors as much as possible when the hint of no error is provided.
    - But, when asking the PARROT-HINT to generate preferred translations, the performance drops considerably.
    - Contrastive instruction only indicates that two translations may have quality differences but not states why, which is difficult for LLMs to identify by themselves.
    - Applying PARROT framework on LLMs based on BLOOMZ-7b-mt perform much better on Zh⇒En and En⇒Zh directions than the LLMs based on LLaMA-7b, which mainly results from the better modeling of Chinese during the pretraining of BLOOMZ-7b-mt.
- **Parameter efficient finetuning may prevent LLMs from overfitting:**
    - Alpaca-LoRA outperforms its full model counterpart.
    - LoRA can prevent LLMs from overfitting the small Alpaca multi-task dataset, leading to a stronger generalization ability.

- Applying LoRA to PARROT exhibits distinct behaviors for high-resource and low-resource translation directions. Specifically, PARROT-HINT-LoRA outperforms the corresponding full model PARROT-HINT on De⇒En but performs much worse on the other directions.

**ANALYSIS:**

- **Effectiveness of Error Guided Instructions:(Parrot Hint)**

Table 4: Effects of error levels as hints during inference on Flores subsets.

| Hint | En⇒De | | Zh⇒En | |
|---|---|---|---|---|
| | BLEU | COMET | BLEU | COMET |
| None | 30.83 | 84.38 | 19.22 | 83.96 |
| No Err. | 32.10 | 84.94 | 21.57 | 83.72 |
| Minor Err. | 28.84 | 83.60 | 20.67 | 82.14 |
| Major Err. | 28.57 | 82.93 | 19.33 | 80.55 |

Table 5: Examples from Flores Zh⇒En subset when translated with varied error levels.

| | COMET | Example |
|---|---|---|
| SRC | n/a | 例如，在卢瓦尔河谷和莱茵河谷参观城堡，或者乘游轮去游览多瑙河沿岸好玩的城市，或者沿伊利运河乘船游玩。 |
| REF | n/a | For example visiting castles in the Loire Valley, the Rhine valley or taking a cruise to interesting cites on the Danube or boating along the Erie Canal. |
| None | 81.22 | For example, visiting the castles in the Rhine Valley and the Lahn Valley, or taking a cruise along the Rhine River to explore the cities along the river, or taking a boat trip along the Illi River. |
| No Err. | 81.49 | For example, visiting the castles in the Rhine Valley and the Ruhr Valley, or taking a cruise to visit the cities along the Rhine River, or taking a boat ride along the Illi River. |
| Minor Err. | 78.12 | For example, visiting the castles in the Rhine Valley and the Ruhr Valley, or taking a cruise to visit the cities along the Rhine River, or taking a boat to play in the cities along the Elbe River. |
| Major Err. | 65.39 | For example, visiting the fortress in the valley of the River Volga and the River Levan, or taking a cruise to visit the cities along the River Volga, or taking a boat to play in the cities along the River Volga. |

It demonstrates that Parrot Hint can place the erroneous translations into other locations of the probability space with the regulation of human annotations. As a result, Parrot Hint is more likely to generate high-quality translation with “no error”.

- **Failure of Contrastive Instruction:**

Table 6: Effects of preference as hint during inference on Flores subsets.

| Hint | En⇒De | | Zh⇒En | |
|---|---|---|---|---|
| | BLEU | COMET | BLEU | COMET |
| None | 30.83 | 84.38 | 19.22 | 83.96 |
| Prefer. | 23.07 | 83.97 | 18.69 | 83.10 |
| Unprefer. | 29.19 | 83.76 | 19.68 | 82.35 |

The unpreferred translations obtain much higher BLUE score though the situation is not for COMET score.

One potential reason is that the WMT systems are so competitive with each other that the quality differences between them are too subtle for the LLM to learn effectively.

**CONCLUSION :**

- Translation instruction, as expected, can improve the translation performance of LLMs significantly, especially for directions from English to other languages.
- Contrastive instruction does not work as expected, which may result from the subtle difference between translations by the competitive WMT systems.
- Error-guided instruction can further improve the performance when asking Parrot to generate translations with no error, indicating the importance of learning from low-quality translations annotated by human.
- Parameter efficient finetuning with low rank adaptation can prevent LLMs from overfitting, which achieves better performance on dominant languages but slows down the learning from other languages.
- With the Alpaca multi-task dataset involved, Parrot can also preserve the capability of general tasks, such as question answering and code generation.