



ParroT: Translating during Chat using Large Language Models tuned with Human Translation and Feedback

Wenxiang Jiao^{1*}, Jen-tse Huang^{1,2}, Wenxuan Wang^{1,2}, Zhiwei He^{1,3}, Tian Liang^{1,4},
Xing Wang¹, Shuming Shi¹, Zhaopeng Tu¹

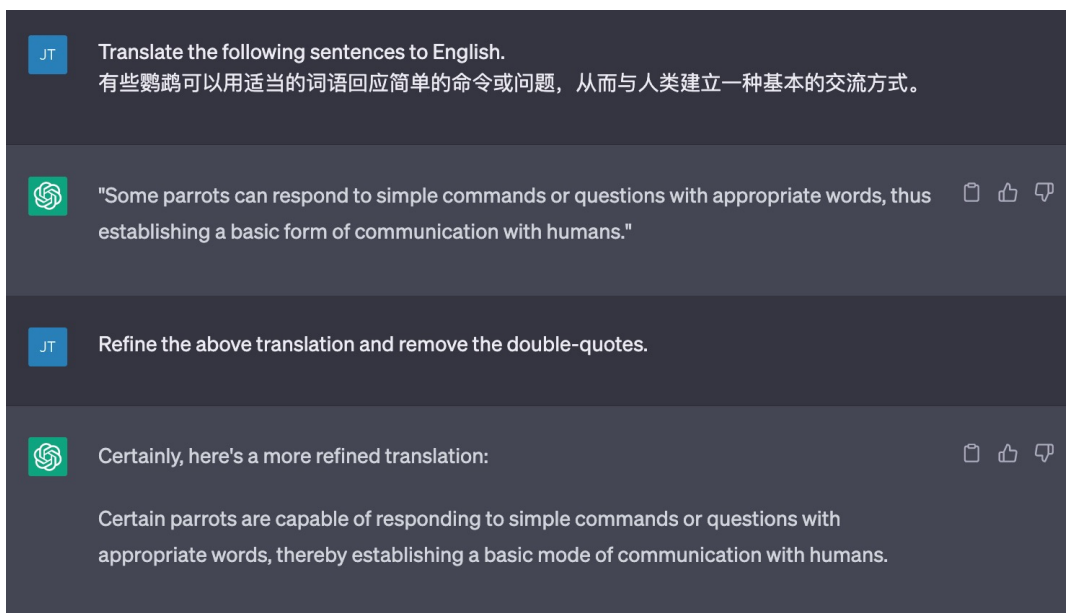
¹Tencent AI Lab

²The Chinese University of Hong Kong

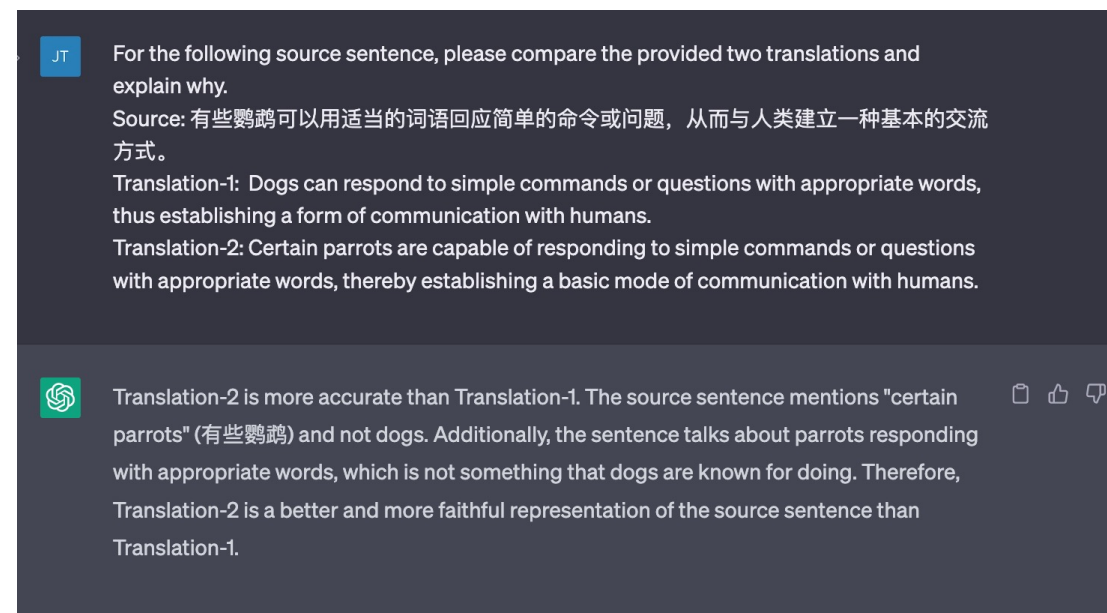
³Shanghai Jiao Tong University

⁴Tsinghua Shenzhen International Graduate School

- Machine translation includes various **sub-tasks** (QE, APE), handled by separate models
- ChatGPT integrates different NLP tasks, enabling **flexible switching** among tasks
- How to explore various translation abilities upon **open-source LLMs**?



Automatic Post-Editing



Quality Comparison

- Leverage human-written translation and feedback data
- Reformulate translation data into **instruction-following** style
- A **Hint** field to incorporate extra requirements



Translate the following sentences to English.

有些鹦鹉可以用适当的词语回应简单的命令或问题，从而与人类建立一种基本的交流方式。



Hint: preferred/no errors/entity alignment/...

Some parrots can respond to simple commands or questions with the appropriate words, establishing a basic form of communication with humans.



Figure 1: Framework of PARROT. Hints are (optional) extra requirements to regulate the translation process.

Parrots are smart birds that can respond to simple commands or questions. The question is whether they're just mimicking, or really intelligent enough to communicate with humans. This is similar to what we currently speculate about LLMs.

➤ **Translation Instruction:** asks LLMs to generate translations based on source sentences

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	<div>### Instruction: \nTranslate the following sentences from Chinese to English.\n\n</div> <div>### Input: \n 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，市场秩序总体平稳。 \n\n</div> <div>### 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.</div>

➤ **Contrastive Instruction:** asks LLMs to generate the translations of two different systems with the preferred one at first

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
Contrastive	<div>### Instruction: \nTranslate the following sentences from Chinese to English.\n\n</div> <div>### Input: \n 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，市场秩序总体平稳。 \n\n</div> <div>### Hint: We prefer to translate it to\n\n</div> <div>### Response: <p>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.</p> rather than <p>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.</p></div>

➤ **Error-Guided Instruction:** asks LLMs to generate the translations with human-annotated errors as the hint

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
Error-Guided	<div>### Instruction: \nTranslate the following sentences from Chinese to English.\n\n</div> <div>### Input: \n 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，市场秩序总体平稳。 \n\n</div> <div>### Hint: A translation with major accuracy/mistranslation errors could be\n\n</div> <div>### Response: The results of the inspection indicate the sufficient supply of living necessities <v>on market-ing</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.</div> <div>### Hint: A translation with minor fluency/grammar errors could be\n\n</div> <div>### 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.</div>

➤ Training Data

- Alpaca Data: general tasks
- [WMT Validation Data](#): human-written translation
- [Multidimensional Quality Metrics](#): human-annotated errors
- Automatically Assessed Data: quality score by COMET metric

➤ Model Training

- Base LLMs: [LLaMA](#)-7b/13b, [BLOOM](#)-560m/7b1
- Model Variants: Alpaca, Parrot-T, Parrot, Parrot-LoRA
- Training Info: DeepSpeed ZeRO 3, 8 Nvidia A100 GPUs

➤ Evaluation

- Test Data: Flores test sets, WMT22 test sets
- Metrics: SacreBLEU, COMET

➤ **LLaMA**: 1) Translation instruction improves translation performance significantly

System	De⇒En		En⇒De		Zh⇒En		En⇒Zh	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Flores Subsets								
Google	45.0	88.7	41.1	88.6	31.6	87.7	43.5	88.4
DeepL	49.2	89.7	41.4	89.0	31.2	87.3	44.3	88.1
ChatGPT	43.7	89.1	38.8	88.1	24.7	85.8	38.2	86.9
GPT-4	46.0	89.3	45.7	89.2	28.5	87.4	42.5	88.4
<i>Base Model: LLaMA-7b</i>								
Vanilla	3.4	60.1	2.4	49.0	1.8	53.7	0.1	47.6
Alpaca	36.6	86.8	23.3	80.5	15.1	81.2	9.8	58.6
Alpaca-LoRA	40.7	87.7	24.6	84.0	16.4	81.5	14.5	70.5
Parrot-T	41.3	87.7	28.5	83.3	19.5	83.1	24.7	79.9
Parrot	41.0	87.9	30.8	84.3	19.2	83.9	25.8	80.1
+ Infer w/ Prefer.	38.1	87.6	23.0	83.9	18.6	83.1	22.5	80.1
+ Infer w/ No Err.	42.2	88.7	32.1	84.9	21.5	83.7	27.4	81.8
Parrot-LoRA	43.8	88.3	29.0	84.9	16.9	80.6	14.8	71.5
+ Infer w/ No Err.	42.0	88.0	29.8	85.4	17.4	81.3	19.8	76.7

➤ LLaMA: 2) Error-guided instruction brings further improvement

System	De⇒En		En⇒De		Zh⇒En		En⇒Zh	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Flores Subsets								
Google	45.0	88.7	41.1	88.6	31.6	87.7	43.5	88.4
DeepL	49.2	89.7	41.4	89.0	31.2	87.3	44.3	88.1
ChatGPT	43.7	89.1	38.8	88.1	24.7	85.8	38.2	86.9
GPT-4	46.0	89.3	45.7	89.2	28.5	87.4	42.5	88.4
<i>Base Model: LLaMA-7b</i>								
Vanilla	3.4	60.1	2.4	49.0	1.8	53.7	0.1	47.6
Alpaca	36.6	86.8	23.3	80.5	15.1	81.2	9.8	58.6
Alpaca-LoRA	40.7	87.7	24.6	84.0	16.4	81.5	14.5	70.5
Parrot-T	41.3	87.7	28.5	83.3	19.5	83.1	24.7	79.9
Parrot	41.0	87.9	30.8	84.3	19.2	83.9	25.8	80.1
+ Infer w/ Prefer.	38.1	87.6	23.0	83.9	18.6	83.1	22.5	80.1
+ Infer w/ No Err.	42.2	88.7	32.1	84.9	21.5	83.7	27.4	81.8
Parrot-LoRA	43.8	88.3	29.0	84.9	16.9	80.6	14.8	71.5
+ Infer w/ No Err.	42.0	88.0	29.8	85.4	17.4	81.3	19.8	76.7

➤ LLaMA: 3) LoRA hinders the learning of other languages

System	De⇒En		En⇒De		Zh⇒En		En⇒Zh	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Flores Subsets								
Google	45.0	88.7	41.1	88.6	31.6	87.7	43.5	88.4
DeepL	49.2	89.7	41.4	89.0	31.2	87.3	44.3	88.1
ChatGPT	43.7	89.1	38.8	88.1	24.7	85.8	38.2	86.9
GPT-4	46.0	89.3	45.7	89.2	28.5	87.4	42.5	88.4
<i>Base Model: LLaMA-7b</i>								
Vanilla	3.4	60.1	2.4	49.0	1.8	53.7	0.1	47.6
Alpaca	36.6	86.8	23.3	80.5	15.1	81.2	9.8	58.6
Alpaca-LoRA	40.7	87.7	24.6	84.0	16.4	81.5	14.5	70.5
Parrot-T	41.3	87.7	28.5	83.3	19.5	83.1	24.7	79.9
Parrot	41.0	87.9	30.8	84.3	19.2	83.9	25.8	80.1
+ Infer w/ Prefer.	38.1	87.6	23.0	83.9	18.6	83.1	22.5	80.1
+ Infer w/ No Err.	42.2	88.7	32.1	84.9	21.5	83.7	27.4	81.8
Parrot-LoRA	43.8	88.3	29.0	84.9	16.9	80.6	14.8	71.5
+ Infer w/ No Err.	42.0	88.0	29.8	85.4	17.4	81.3	19.8	76.7

➤ LLaMA: 4) Larger models show stronger potential

System	De⇒En		En⇒De		Zh⇒En		En⇒Zh	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
WMT22 Test Sets								
Google	33.3	84.8	38.4	87.1	28.6	80.9	49.9	87.4
DeepL	32.8	84.7	36.2	87.9	24.2	79.3	44.5	86.4
GPT-4	33.4	84.9	34.5	87.4	24.8	82.3	41.3	87.0
<i>Base Model: LLaMA-7b</i>								
Vanilla	2.9	52.8	1.6	45.3	1.2	50.3	0.3	46.3
Alpaca	27.8	82.3	20.1	78.1	14.2	74.0	10.4	62.1
Alpaca-LoRA	28.9	83.2	22.1	81.3	16.1	75.6	16.3	70.6
ParroT-T	26.6	82.5	24.0	80.4	18.1	75.3	27.0	78.4
ParroT	27.3	82.4	24.6	81.2	18.9	75.2	28.1	79.3
+ Infer w/ No Err.	27.3	82.4	26.1	81.6	20.2	75.9	30.3	80.3
ParroT-LoRA	28.8	82.8	24.0	81.4	18.2	74.7	19.9	73.7
+ Infer w/ No Err.	29.8	83.0	24.8	81.6	19.2	75.0	20.7	74.5
<i>Base Model: LLaMA-13b</i>								
Alpaca	29.7	83.1	21.4	79.4	16.2	75.9	17.6	70.8
ParroT	27.6	83.2	27.0	82.8	19.9	75.8	30.9	81.1
+ Infer w/ No Err.	31.1	83.6	28.1	82.6	21.7	76.7	31.7	81.0

➤ **BLOOM**: 1) Significant improvement; 2) Better than LLaMA on ZH-EN

Table 4: Translation performance of BLOOM models on WMT22 test sets.

System	De⇒En		En⇒De		Zh⇒En		En⇒Zh	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
<i>Base Model: BLOOMZ-560m</i>								
Alpaca	4.4	55.2	0.5	30.8	6.9	70.1	2.0	54.0
ParroT	16.4	68.9	13.3	57.7	16.0	74.8	25.4	79.0
+ Infer w/ No Err.	16.9	69.3	12.8	56.8	15.7	75.0	26.3	79.5
<i>Base Model: BLOOMZ-7b1-mt</i>								
Alpaca	17.6	73.0	3.1	44.5	13.0	76.4	23.9	81.8
ParroT	23.1	77.6	20.0	72.7	21.4	78.5	32.4	83.6
+ Infer w/ No Err.	24.9	78.0	20.5	73.6	22.7	79.0	34.5	83.5

➤ Analysis on EGI: Controllable generation with error levels

- None: instruct_inf.txt
 - Translate the following sentences from [SRC] to [TGT].
- No Errors: instruct_inf_e2t.txt
 - Translate the following sentences from [SRC] to [TGT].###A translation with no errors could be
- Minor Errors: instruct_inf_e2t_minor.txt
 - Translate the following sentences from [SRC] to [TGT].###A translation with minor errors could be
- Major Errors: instruct_inf_e2t_major.txt
 - Translate the following sentences from [SRC] to [TGT].###A translation with major errors could be

Table 5: Effects of error levels as hints during inference.

Red : improvement; **Green** : degradation.

Hint	En⇒De		Zh⇒En	
	BLEU	COMET	BLEU	COMET
None	30.8	84.3	19.2	83.9
No Err.	32.1	84.9	21.5	83.7
Minor Err.	28.8	83.6	20.6	82.1
Major Err.	28.5	82.9	19.3	80.5

➤ Analysis on EGI: Controllable generation with error levels

Table 6: Examples from Flores Zh⇒En subset when translated with varied error levels. **Red**: translation errors.

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	81.2	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.	81.4	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.	78.1	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.	65.3	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 .

wxjiao / ParroT

Q Type / to search

>

+

<> Code

Issues 3

Pull requests


Actions

Projects

Security 103

Insights

Settings

 **ParroT** Public

Unpin

Unwatch 2

Fork 18


Star 148

master 1 branch 3 tags

Go to file

Add file

<> Code

 wxjiao Update README.md 706fea0 on Oct 12 177 commits

data	added alpaca-gpt4 data	6 months ago
docs	backup README	6 months ago
scripts	scripts for data conversion	7 months ago
test	updated test sets	6 months ago
train	print info for loading lora weights in inference_lora.py	3 months ago
transformers	fixed data streaming for local large datasets	2 months ago
README.md	Update README.md	last month
requirements.txt	added flash-attention for BLOOM	4 months ago

About

The ParroT framework to enhance and regulate the Translation Abilities during Chat based on open-sourced LLMs (e.g., LLaMA-7b, Bloomz-7b1-mt) and human written translation and evaluation data.

machine-translation

llama

lora

contrastive

gpt-4

chatgpt

human-feedback

instruction-tuning

bloomz

error-guided

Readme

Activity

148 stars

Thanks!