天津大学
Tianjin University

Can Large Language Models Learn Translation Robustness from Noisy-Source In-context Demonstraions?

Leiyu Pan, Yongqi Leng, Deyi Xiong

College of Intelligence and Computing, Tianjin University, Tianjin, China

https://github.com/tjunlp-lab/llm_translate_robust

TJU NLP

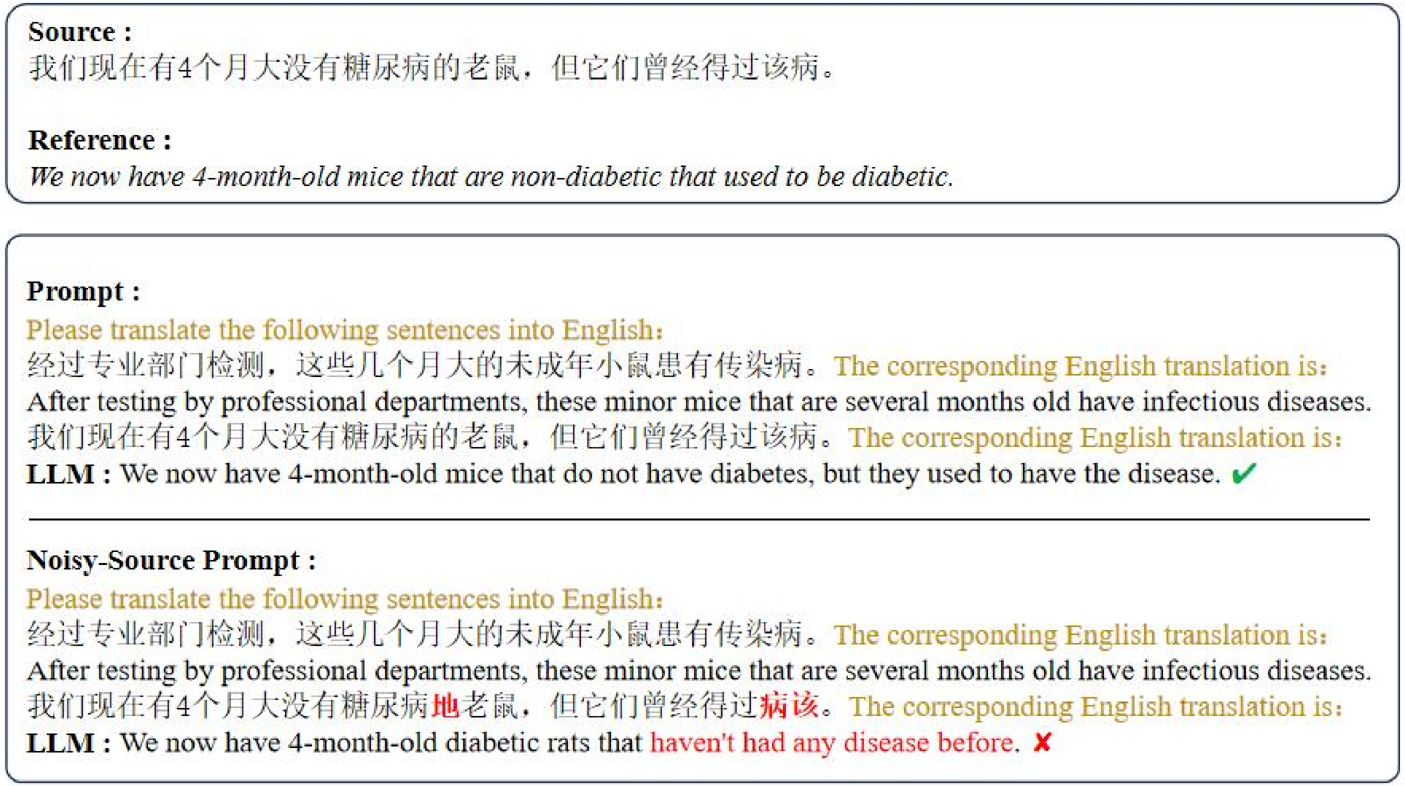
the 2024 Joint International Conference on Computational Linguistics,
Language Resources and Evaluation (LREC-COLING 2024)

Background

- Machine Translation with Large Language Models (LLMs)
  - Direct translation using LLMs
  - In-context learning for translation using LLMs
- Robust Machine Translation
  - Translation of sentences containing noise

Challenge

- Robust Machine Translation in LLMs
  - Limitations of LLMs in translating noisy sentences (even use in-context learning)



Source :
我们现在有4个月大没有糖尿病的老鼠，但它们曾经得过该病。

Reference :
We now have 4-month-old mice that are non-diabetic that used to be diabetic.

Prompt :
Please translate the following sentences into English：
经过专业部门检测，这些几个月大的未成年小鼠患有传染病。The corresponding English translation is：
After testing by professional departments, these minor mice that are several months old have infectious diseases.
我们现在有4个月大没有糖尿病的老鼠，但它们曾经得过该病。The corresponding English translation is：
LLM : We now have 4-month-old mice that do not have diabetes, but they used to have the disease. ✔

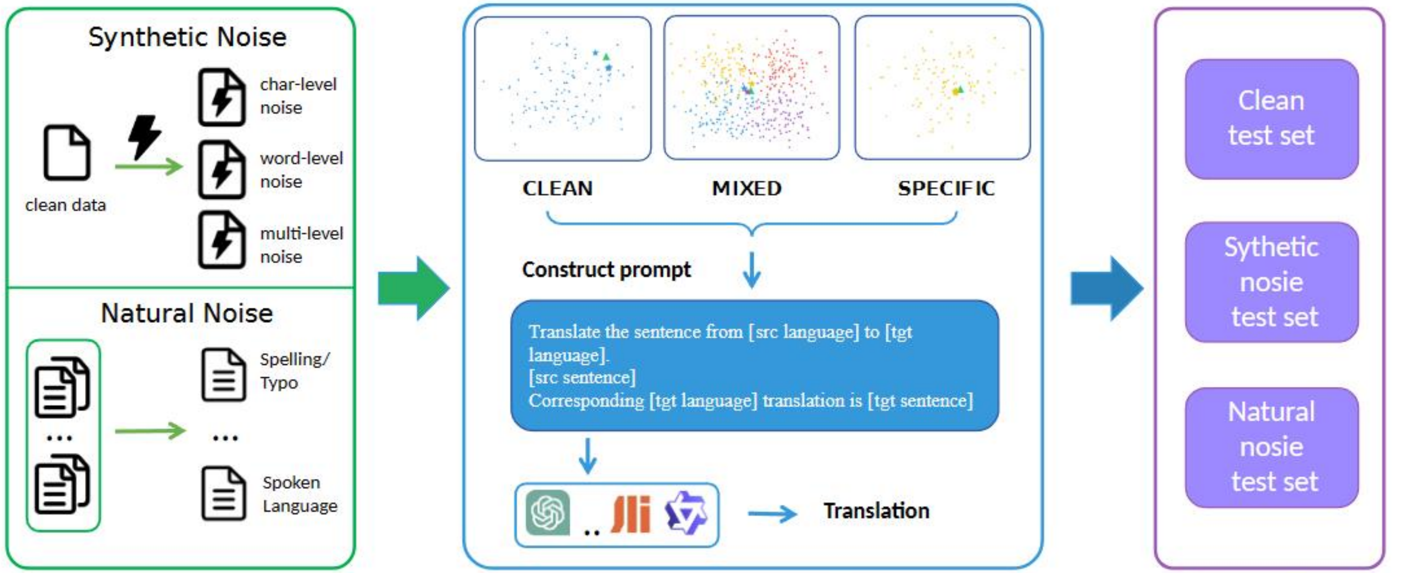
Noisy-Source Prompt :
Please translate the following sentences into English：
经过专业部门检测，这些几个月大的未成年小鼠患有传染病。The corresponding English translation is：
After testing by professional departments, these minor mice that are several months old have infectious diseases.
我们现在有4个月大没有糖尿病**地**老鼠，但它们曾经得过**病该**。The corresponding English translation is：
LLM : We now have 4-month-old diabetic rats that haven't had any disease before. ✘

Motivation

- In-context learning allows LLMs to learn from in-context demonstrations and thus accomplish the tasks better than using LLMs directly.
- LLMs benefit more from in-context demonstrations that are more similar to the test sample.
- Issues that this work seeks to explore:
  - Can LLMs learn translation robustness from noisy-source in-context demonstrations?
  - Whether LLMs are more likely to learn translation robustness from in-context demonstrations that are consistent with the type of noise in the sample to be translated?

Approach

- propose a research scheme for investigating the translation robustness of LLMs
  - noisy data preparation
  - in-context learning for translation using LLMs
  - evaluation of translation robustness



Translate the sentence from [src language] to [tgt language].
[src sentence]
Corresponding [tgt language] translation is [tgt sentence]

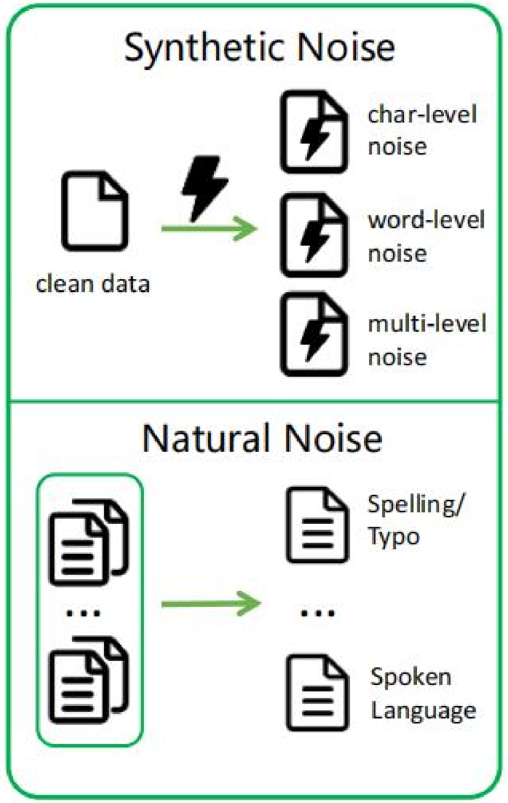
Noisy Data Preparation

In-context Learning for Translation using LLMs

Evaluation of Translation Robustness

Approach

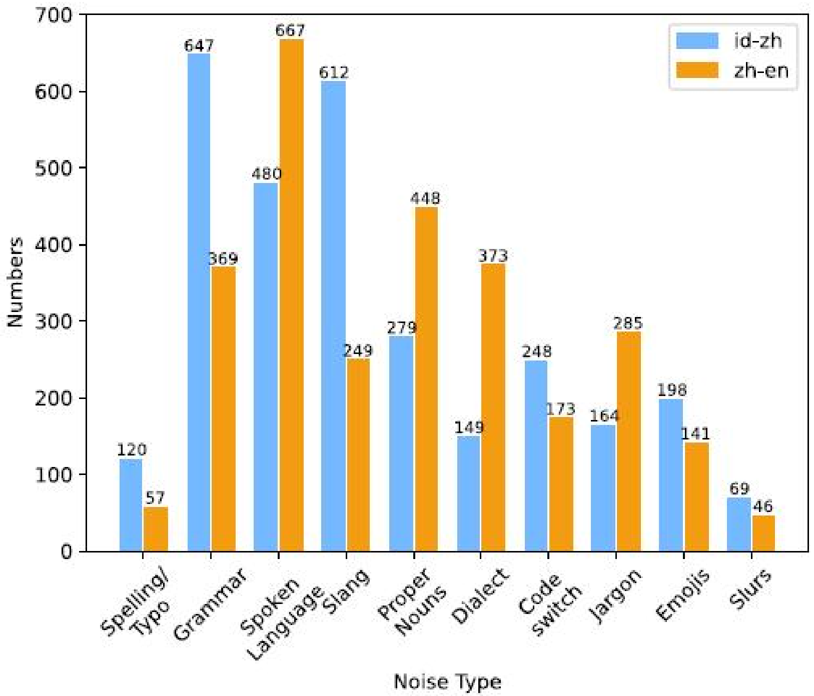
- Noisy Data Preparation
  - Parallel corpus of high- and low-resource
  - **Synthetic noise**: character-level, word-level, multi-level noise
  - **Natural noise**: spelling/typographical errors, grammar errors, spoken language, slang, proper nouns, dialects, code switching, jargon, emojis, slurs



Noisy Data Preparation

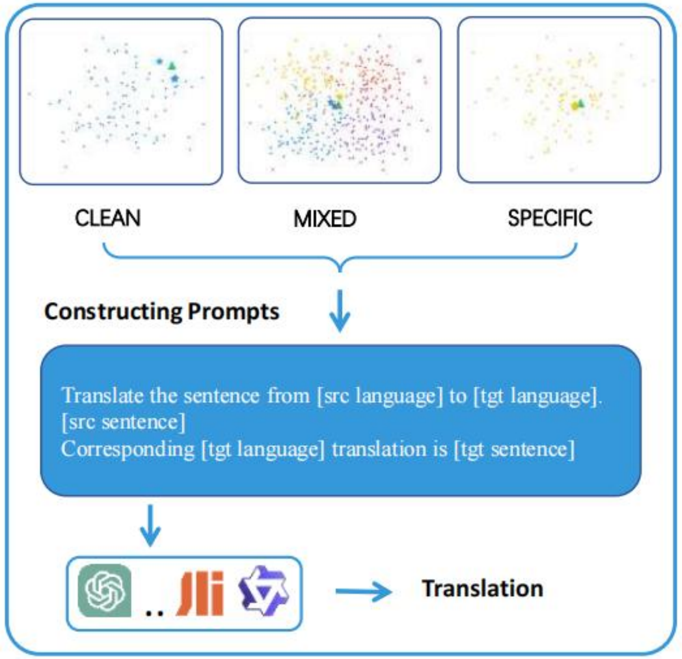
Approach

- Natural Noise Data Preparation
  - **Rule-based labeling**: emoji (regular expression) / dialect (opencc) / code switching (regular expression)
  - **Model-based labeling**: use GPT-3.5 turbo api
  - **Manual labeling**



Approach

- In-context Learning for Translation using LLMs
  - **CLEAN**: sample in-context demonstrations from clean data
  - **MIXED**: sample in-context demonstrations from mixed noise categories of data
  - **SPECIFIC**: sampling in-context demonstrations from data with the same noise type as the sentence to be translated



Translate the sentence from [src language] to [tgt language].
[src sentence]
Corresponding [tgt language] translation is [tgt sentence]

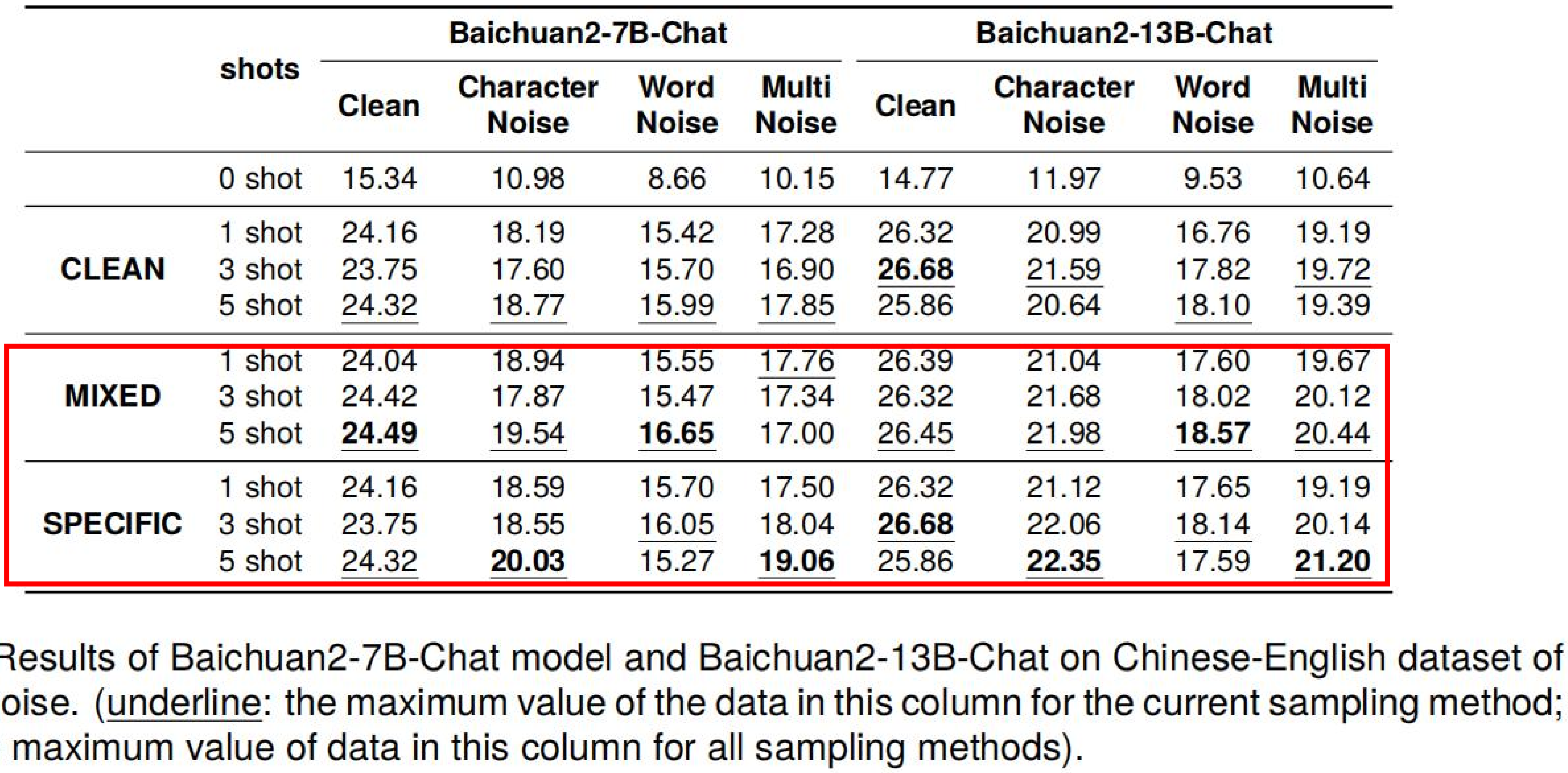
In-context Learning
for Translation using LLMs

Experiment Settings

- Data
  - Synthetic Noise Data
    - WMT News test set & TED TALKS 2020
    - attack settings: 30% attacked words & equal probability of each attack operation
  - Natural Noise Data
    - MMTC dataset & ID-ZH-MTRobustEval
- Model
  - Baichuan2/Qwen/InternLM
- Sample Details
  - determine the sampling set
  - select the most similar in-context demonstrations for sentence embedding

Experiment Results

- LLMs can learn translation robustness from in-context examples with synthetic noise.

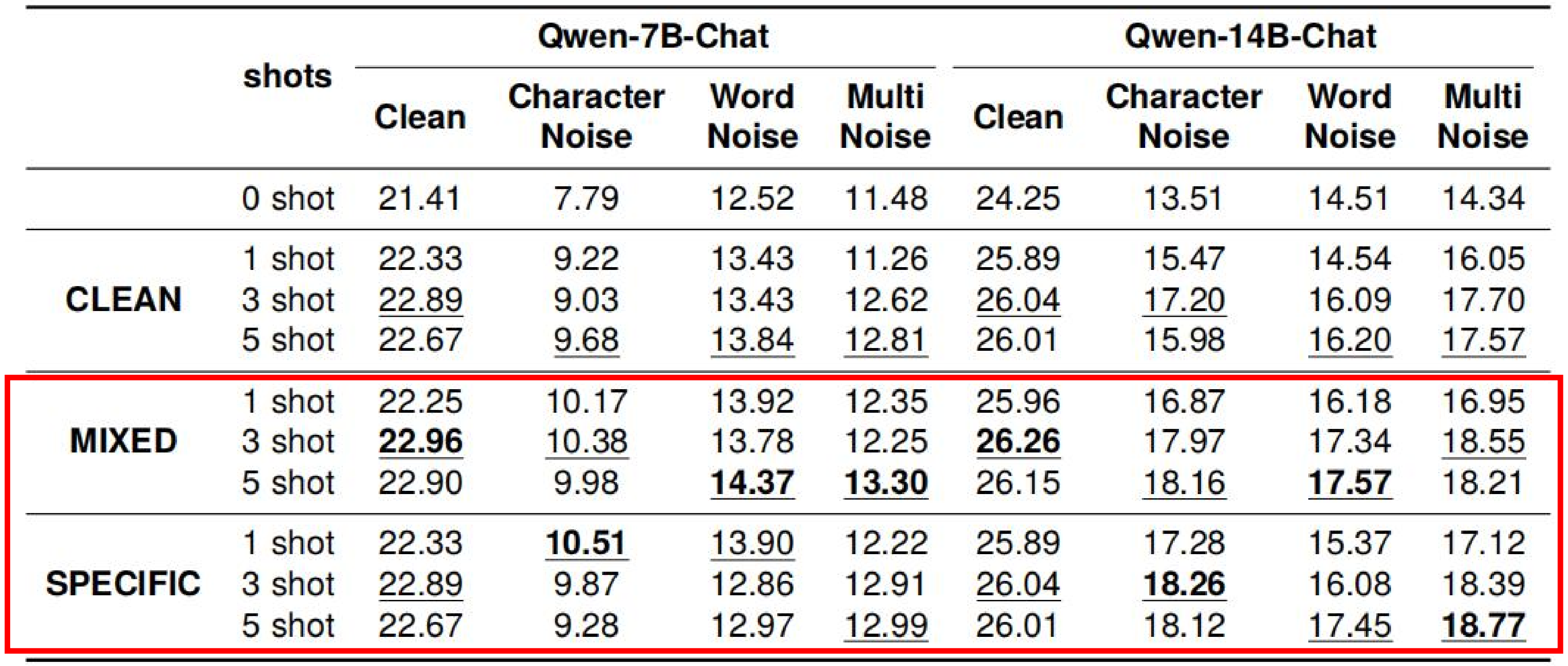


	shots	Baichuan2-7B-Chat				Baichuan2-13B-Chat			
		Clean	Character Noise	Word Noise	Multi Noise	Clean	Character Noise	Word Noise	Multi Noise
	0 shot	15.34	10.98	8.66	10.15	14.77	11.97	9.53	10.64
CLEAN	1 shot	24.16	18.19	15.42	17.28	26.32	20.99	16.76	19.19
	3 shot	23.75	17.60	15.70	16.90	**26.68**	21.59	17.82	19.72
	5 shot	24.32	18.77	15.99	17.85	25.86	20.64	18.10	19.39
MIXED	1 shot	24.04	18.94	15.55	17.76	26.39	21.04	17.60	19.67
	3 shot	24.42	17.87	15.47	17.34	26.32	21.68	18.02	20.12
	5 shot	**24.49**	19.54	**16.65**	17.00	26.45	21.98	**18.57**	20.44
SPECIFIC	1 shot	24.16	18.59	15.70	17.50	26.32	21.12	17.65	19.19
	3 shot	23.75	18.55	16.05	18.04	**26.68**	22.06	18.14	20.14
	5 shot	24.32	**20.03**	15.27	**19.06**	25.86	**22.35**	17.59	**21.20**

Table 1: Results of Baichuan2-7B-Chat model and Baichuan2-13B-Chat on Chinese-English dataset of sythetic noise. (underline: the maximum value of the data in this column for the current sampling method; **bold**: the maximum value of data in this column for all sampling methods).

Experiment Results

- LLMs can learn translation robustness from in-context examples with synthetic noise.

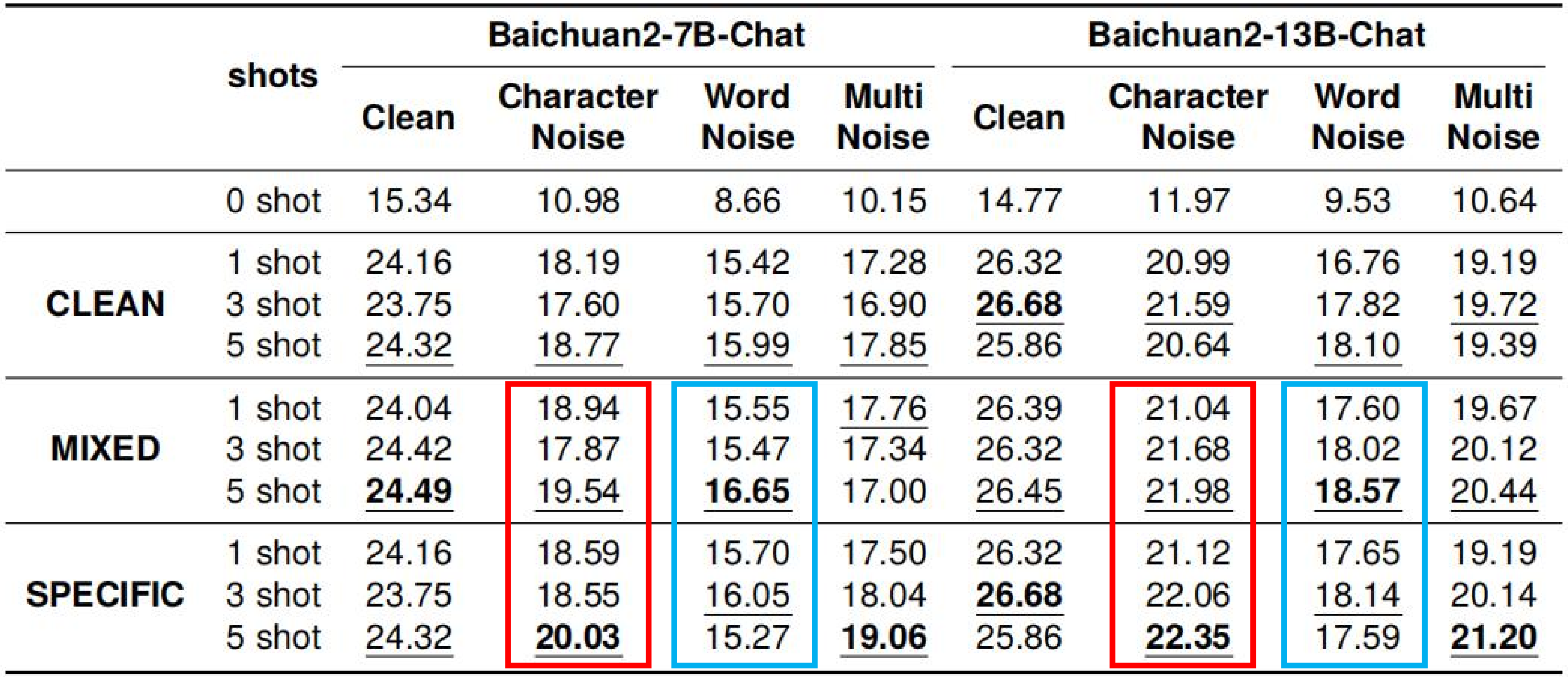


	shots	Qwen-7B-Chat				Qwen-14B-Chat			
		Clean	Character Noise	Word Noise	Multi Noise	Clean	Character Noise	Word Noise	Multi Noise
	0 shot	21.41	7.79	12.52	11.48	24.25	13.51	14.51	14.34
CLEAN	1 shot	22.33	9.22	13.43	11.26	25.89	15.47	14.54	16.05
	3 shot	22.89	9.03	13.43	12.62	26.04	17.20	16.09	17.70
	5 shot	22.67	9.68	13.84	12.81	26.01	15.98	16.20	17.57
MIXED	1 shot	22.25	10.17	13.92	12.35	25.96	16.87	16.18	16.95
	3 shot	**22.96**	10.38	13.78	12.25	**26.26**	17.97	17.34	18.55
	5 shot	22.90	9.98	**14.37**	**13.30**	26.15	18.16	**17.57**	18.21
SPECIFIC	1 shot	22.33	**10.51**	13.90	12.22	25.89	17.28	15.37	17.12
	3 shot	22.89	9.87	12.86	12.91	26.04	**18.26**	16.08	18.39
	5 shot	22.67	9.28	12.97	12.99	26.01	18.12	17.45	**18.77**

Table 2: Results of Qwen-7B-Chat model and Qwen-14B-Chat on Indonesian-Chinese dataset of sythetic noise.

Experiment Results

- LLMs are more likely to learn robustness to character-level noise through type-specific synthetic noise and robustness to word-level noise through mixed-type synthetic noise.

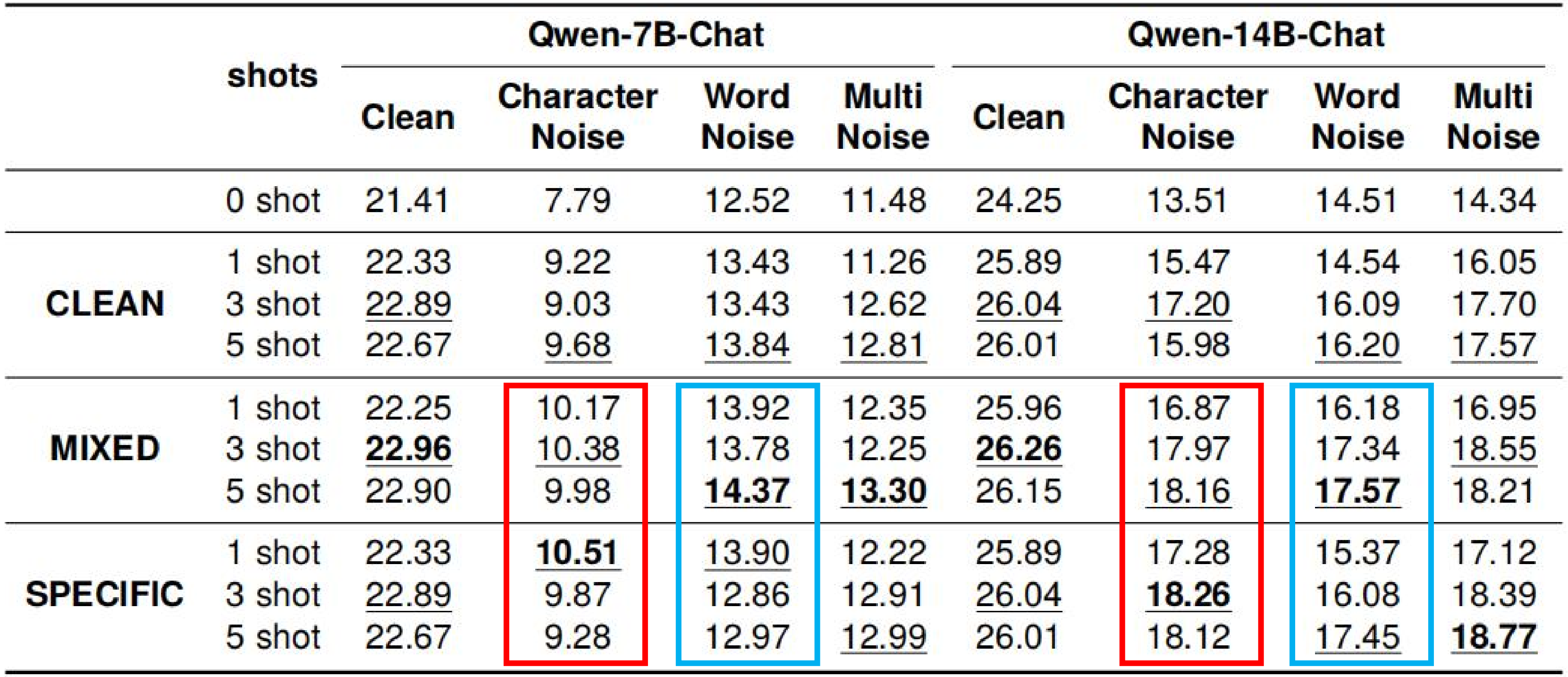


	shots	Baichuan2-7B-Chat				Baichuan2-13B-Chat			
		Clean	Character Noise	Word Noise	Multi Noise	Clean	Character Noise	Word Noise	Multi Noise
	0 shot	15.34	10.98	8.66	10.15	14.77	11.97	9.53	10.64
CLEAN	1 shot	24.16	18.19	15.42	17.28	26.32	20.99	16.76	19.19
	3 shot	23.75	17.60	15.70	16.90	**26.68**	21.59	17.82	19.72
	5 shot	24.32	18.77	15.99	17.85	25.86	20.64	18.10	19.39
MIXED	1 shot	24.04	18.94	15.55	17.76	26.39	21.04	17.60	19.67
	3 shot	24.42	17.87	15.47	17.34	26.32	21.68	18.02	20.12
	5 shot	**24.49**	19.54	**16.65**	17.00	26.45	21.98	**18.57**	20.44
SPECIFIC	1 shot	24.16	18.59	15.70	17.50	26.32	21.12	17.65	19.19
	3 shot	23.75	18.55	16.05	18.04	**26.68**	22.06	18.14	20.14
	5 shot	24.32	**20.03**	15.27	**19.06**	25.86	**22.35**	17.59	**21.20**

Table 1: Results of Baichuan2-7B-Chat model and Baichuan2-13B-Chat on Chinese-English dataset of sythetic noise. (underline: the maximum value of the data in this column for the current sampling method; **bold**: the maximum value of data in this column for all sampling methods).

Experiment Results

- LLMs are more likely to learn robustness to character-level noise through type-specific synthetic noise and robustness to word-level noise through mixed-type synthetic noise.

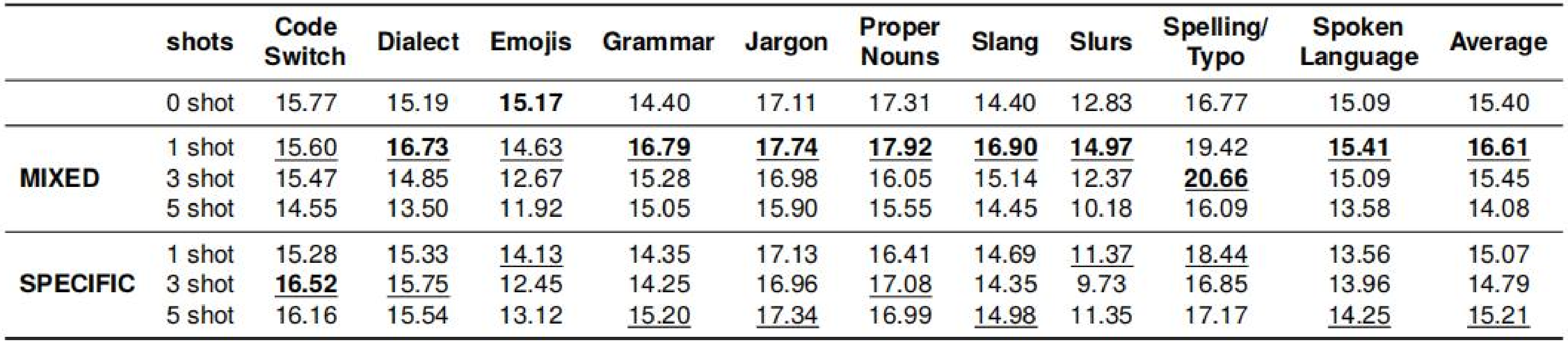


	shots	Qwen-7B-Chat				Qwen-14B-Chat			
		Clean	Character Noise	Word Noise	Multi Noise	Clean	Character Noise	Word Noise	Multi Noise
	0 shot	21.41	7.79	12.52	11.48	24.25	13.51	14.51	14.34
CLEAN	1 shot	22.33	9.22	13.43	11.26	25.89	15.47	14.54	16.05
	3 shot	22.89	9.03	13.43	12.62	26.04	17.20	16.09	17.70
	5 shot	22.67	9.68	13.84	12.81	26.01	15.98	16.20	17.57
MIXED	1 shot	22.25	10.17	13.92	12.35	25.96	16.87	16.18	16.95
	3 shot	**22.96**	10.38	13.78	12.25	**26.26**	17.97	17.34	18.55
	5 shot	22.90	9.98	**14.37**	**13.30**	26.15	18.16	**17.57**	18.21
SPECIFIC	1 shot	22.33	**10.51**	13.90	12.22	25.89	17.28	15.37	17.12
	3 shot	22.89	9.87	12.86	12.91	26.04	**18.26**	16.08	18.39
	5 shot	22.67	9.28	12.97	12.99	26.01	18.12	17.45	**18.77**

Table 2: Results of Qwen-7B-Chat model and Qwen-14B-Chat on Indonesian-Chinese dataset of sythetic noise.

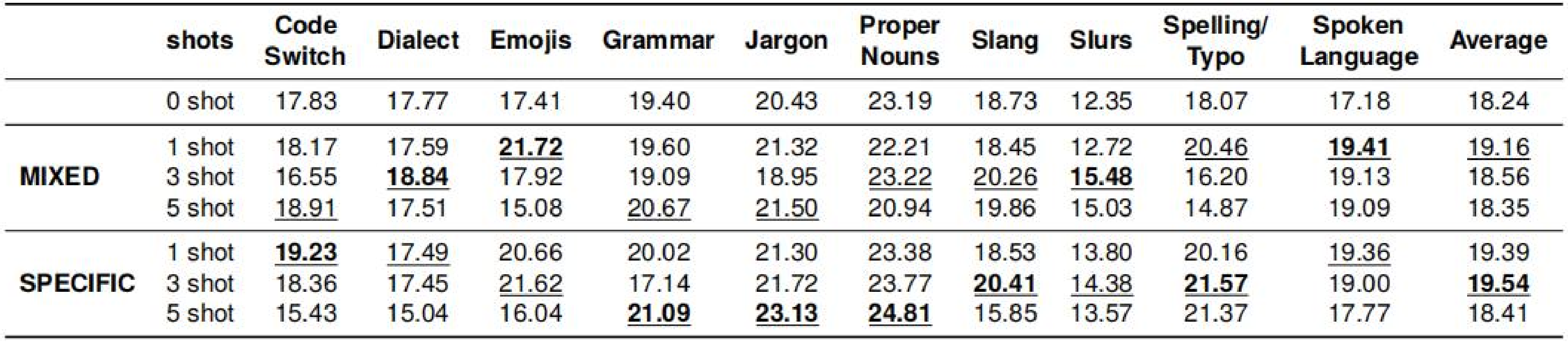
Experiment Results

- The robustness of LLMs in learning from various types of natural noises varies across high and low-resource languages.



	shots	Code Switch	Dialect	Emojis	Grammar	Jargon	Proper Nouns	Slang	Slurs	Spelling/ Typo	Spoken Language	Average
	0 shot	15.77	15.19	**15.17**	14.40	17.11	17.31	14.40	12.83	16.77	15.09	15.40
MIXED	1 shot	15.60	**16.73**	14.63	**16.79**	**17.74**	**17.92**	**16.90**	**14.97**	19.42	**15.41**	**16.61**
	3 shot	15.47	14.85	12.67	15.28	16.98	16.05	15.14	12.37	**20.66**	15.09	15.45
	5 shot	14.55	13.50	11.92	15.05	15.90	15.55	14.45	10.18	16.09	13.58	14.08
SPECIFIC	1 shot	15.28	15.33	14.13	14.35	17.13	16.41	14.69	11.37	18.44	13.56	15.07
	3 shot	**16.52**	15.75	12.45	14.25	16.96	17.08	14.35	9.73	16.85	13.96	14.79
	5 shot	16.16	15.54	13.12	15.20	17.34	16.99	14.98	11.35	17.17	14.25	15.21

Table 3: Results of Baichuan2-7B-Chat model on Chinese-English dataset of natural noise.



	shots	Code Switch	Dialect	Emojis	Grammar	Jargon	Proper Nouns	Slang	Slurs	Spelling/ Typo	Spoken Language	Average
	0 shot	17.83	17.77	17.41	19.40	20.43	23.19	18.73	12.35	18.07	17.18	18.24
MIXED	1 shot	18.17	17.59	**21.72**	19.60	21.32	22.21	18.45	12.72	20.46	**19.41**	19.16
	3 shot	16.55	**18.84**	17.92	19.09	18.95	23.22	20.26	**15.48**	16.20	19.13	18.56
	5 shot	18.91	17.51	15.08	20.67	21.50	20.94	19.86	15.03	14.87	19.09	18.35
SPECIFIC	1 shot	**19.23**	17.49	20.66	20.02	21.30	23.38	18.53	13.80	20.16	19.36	19.39
	3 shot	18.36	17.45	21.62	17.14	21.72	23.77	**20.41**	14.38	**21.57**	19.00	**19.54**
	5 shot	15.43	15.04	16.04	**21.09**	**23.13**	**24.81**	15.85	13.57	21.37	17.77	18.41

Table 4: Results of Qwen-7B-Chat model on Indonesian-Chinese dataset of natural noise.

Analysis

- Effect of Noise Ratio in Demonstrations

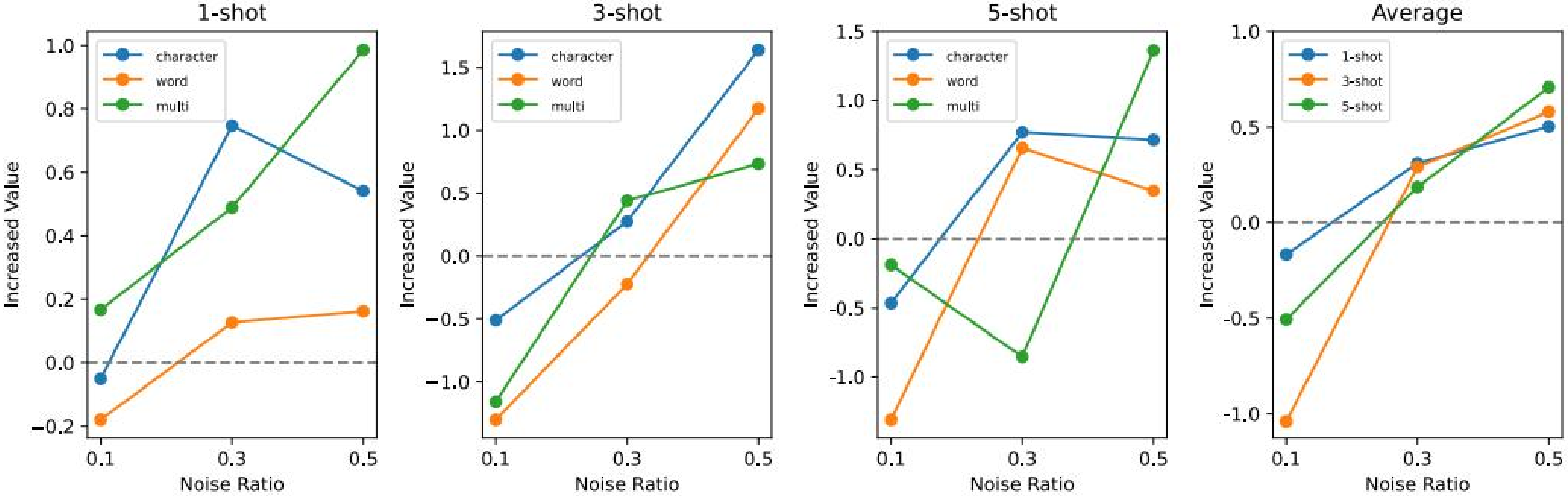
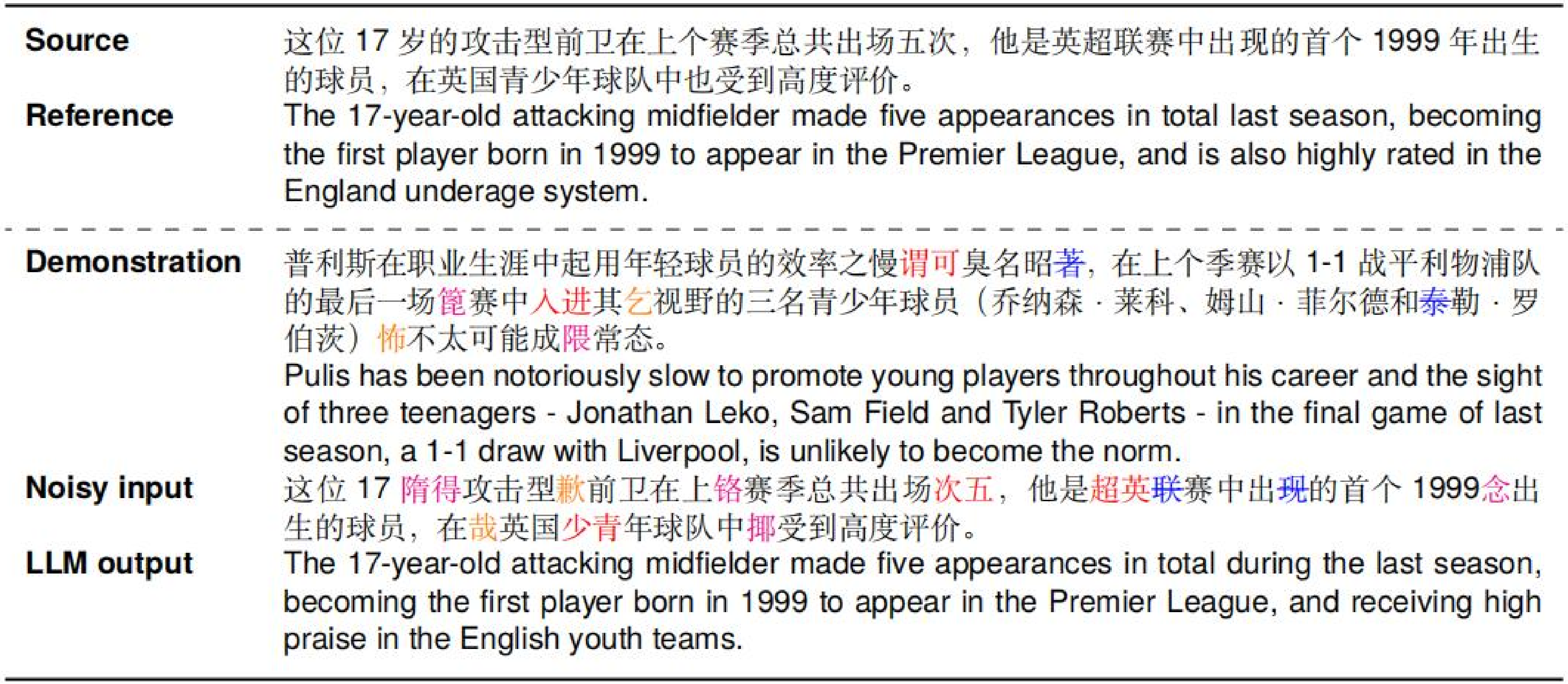


Figure 3: The relationship between the BLEU score change and noise proportion in demonstration examples when comparing the MIXED sampling method with three types of synthetic noise on a Chinese-English dataset to the CLEAN sampling method. The first three subfigures respectively demonstrate the relationship between the noise ratio and the change in BLEU score for 1-shot, 3-shot, and 5-shot settings. The fourth subfigure averages the BLEU score changes for all types of noise under the same shot and noise ratio settings, and displays the relationship between the noise ratio and this average value.

Analysis

- Case Study

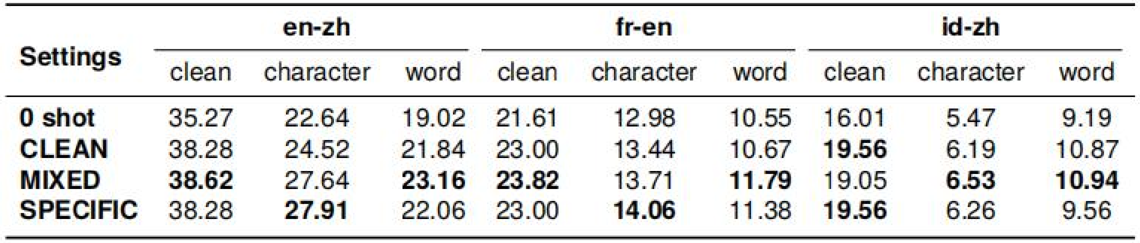


Source	这位 17 岁的攻击型前卫在上个赛季总共出场五次，他是英超联赛中出现的首个 1999 年出生的球员，在英国青少年球队中也受到高度评价。
Reference	The 17-year-old attacking midfielder made five appearances in total last season, becoming the first player born in 1999 to appear in the Premier League, and is also highly rated in the England underage system.
Demonstration	普利斯在职业生涯中起用年轻球员的效率之慢谓可臭名昭~~著~~，在上个季赛以 1-1 战平利物浦队的最后一场篦赛中入进其乞视野的三名青少年球员（乔纳森 · 莱科、姆山 · 菲尔德和~~泰~~勒 · 罗伯茨）怖不太可能成隈常态。 Pulis has been notoriously slow to promote young players throughout his career and the sight of three teenagers - Jonathan Leko, Sam Field and Tyler Roberts - in the final game of last season, a 1-1 draw with Liverpool, is unlikely to become the norm.
Noisy input	这位 17 隋得攻击型歉前卫在上铬赛季总共出场次五，他是超英~~联~~赛中出~~现~~的首个 1999念出生的球员，在哉英国少青年球队中揶受到高度评价。
LLM output	The 17-year-old attacking midfielder made five appearances in total during the last season, becoming the first player born in 1999 to appear in the Premier League, and receiving high praise in the English youth teams.

Table 5: LLM for generating cases in the Chinese-English dataset, where demonstration examples are selected for 1 shot. (Orange: Insertion noise; Red: Swapping noise; Magenta: Replacement noise; ~~Blue~~: Deletion noise.)

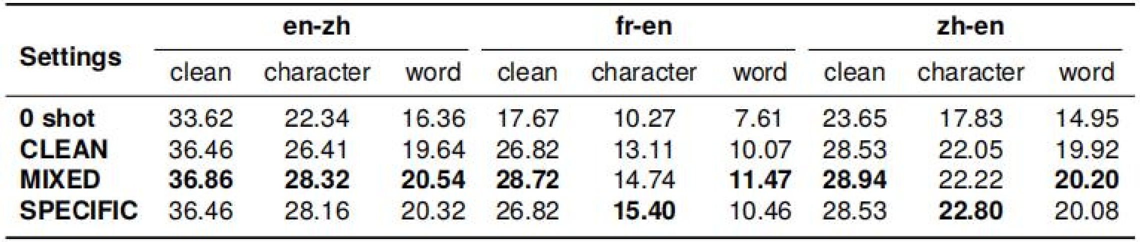
Analysis

- Results for Other Models and Language Pairs



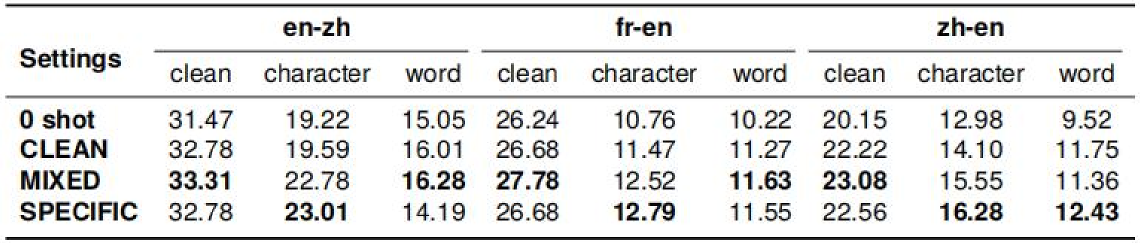
Settings	en-zh			fr-en			id-zh		
	clean	character	word	clean	character	word	clean	character	word
0 shot	35.27	22.64	19.02	21.61	12.98	10.55	16.01	5.47	9.19
CLEAN	38.28	24.52	21.84	23.00	13.44	10.67	**19.56**	6.19	10.87
MIXED	**38.62**	27.64	**23.16**	**23.82**	13.71	**11.79**	19.05	**6.53**	**10.94**
SPECIFIC	38.28	**27.91**	22.06	23.00	**14.06**	11.38	**19.56**	6.26	9.56

Table 6: Results of Baichuan2-7B-Chat on the synthetic noise dataset under various settings.



Settings	en-zh			fr-en			zh-en		
	clean	character	word	clean	character	word	clean	character	word
0 shot	33.62	22.34	16.36	17.67	10.27	7.61	23.65	17.83	14.95
CLEAN	36.46	26.41	19.64	26.82	13.11	10.07	28.53	22.05	19.92
MIXED	**36.86**	**28.32**	**20.54**	**28.72**	14.74	**11.47**	**28.94**	22.22	**20.20**
SPECIFIC	36.46	28.16	20.32	26.82	**15.40**	10.46	28.53	**22.80**	20.08

Table 7: Results of Qwen-7B-Chat on the synthetic noise dataset under various settings.



Settings	en-zh			fr-en			zh-en		
	clean	character	word	clean	character	word	clean	character	word
0 shot	31.47	19.22	15.05	26.24	10.76	10.22	20.15	12.98	9.52
CLEAN	32.78	19.59	16.01	26.68	11.47	11.27	22.22	14.10	11.75
MIXED	**33.31**	22.78	**16.28**	**27.78**	12.52	**11.63**	**23.08**	15.55	11.36
SPECIFIC	32.78	**23.01**	14.19	26.68	**12.79**	11.55	22.56	**16.28**	**12.43**

Table 8: Results of InternLM-Chat-7B on the synthetic noise dataset under various settings.

Thank you!