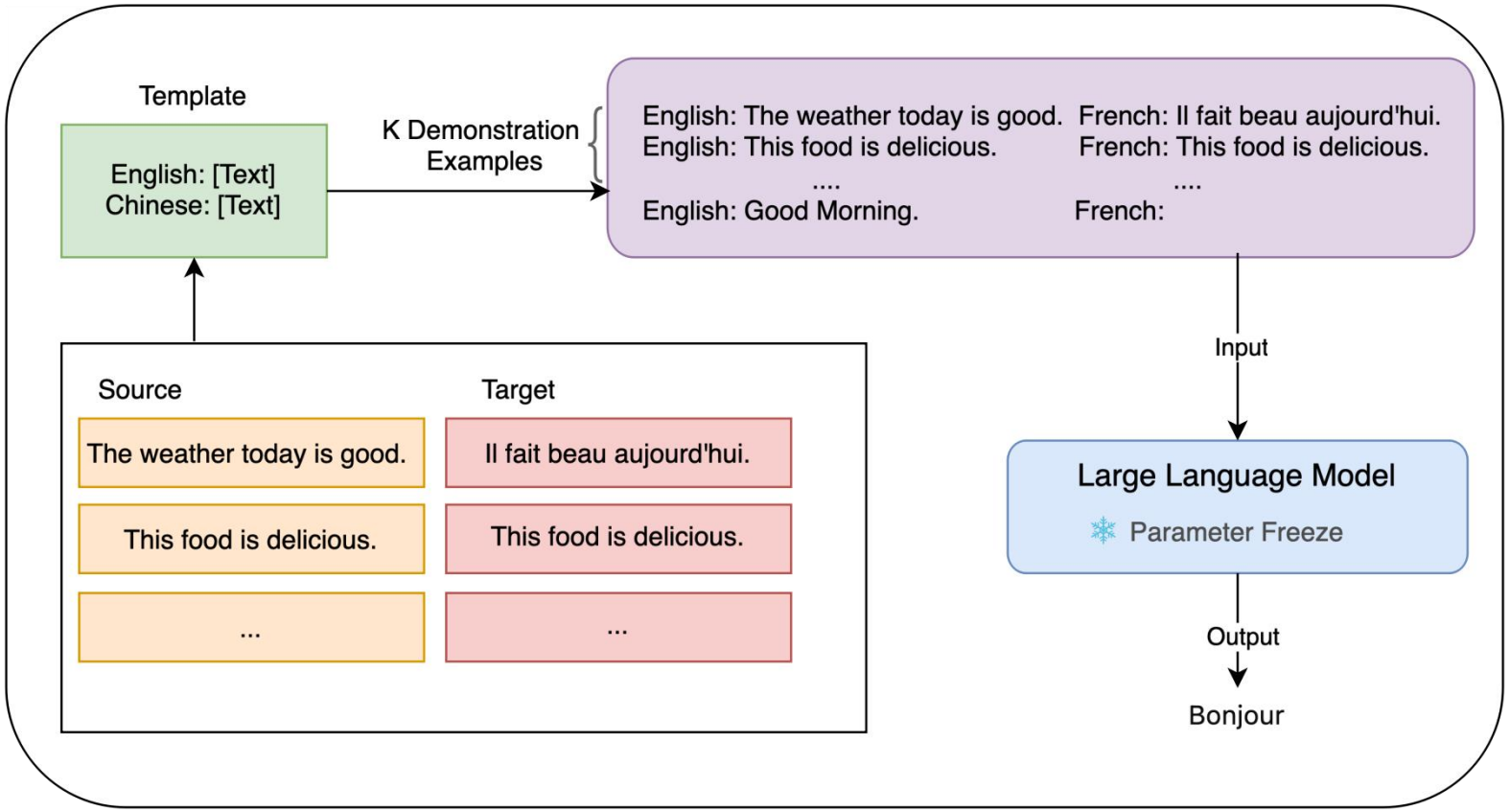
天津大學
Tianjin University

Towards Robust In-Context Learning for Machine Translation with Large Language Models

CONTENTS

- Background
- Preliminary Study
- Method
- Experiments

BACKGROUND



In-Context Learning Paradigm

PRELIMINARY STUDY

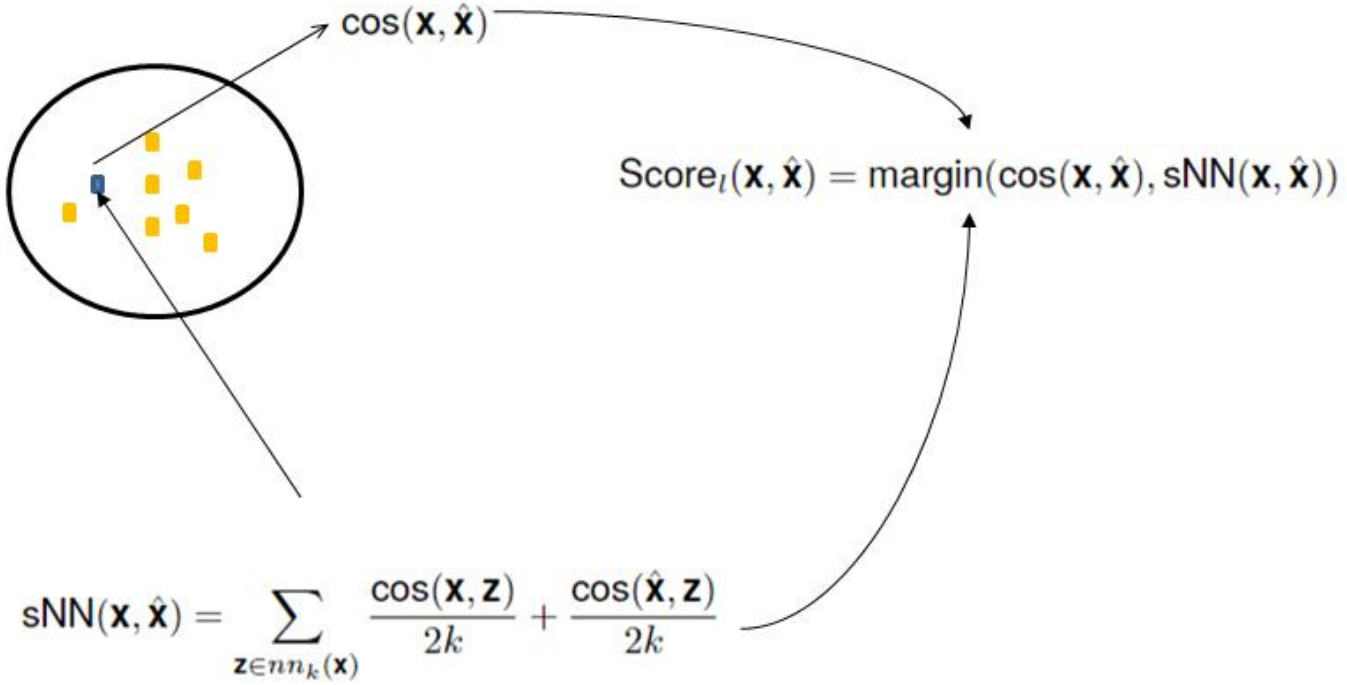
Similarity score		> 0.9	0.75 ~ 0.89	0.6 ~ 0.74	<0.6
en-fr	1-shot	65	330	1123	482
	3-shot	29	139	976	856
	5-shot	23	89	828	1060
en-es	1-shot	108	447	1007	438
	3-shot	42	226	969	763
	5-shot	31	150	888	931

Language & context		>0.9	0.75~ 0.89	0.6~ 0.74	<0.6
en-fr	1-shot	38.73	30.47	22.43	19.70
	3-shot	41.42	29.37	22.69	18.51
	5-shot	45.14	28.75	21.74	18.00
en-es	1-shot	44.82	29.59	23.90	20.25
	3-shot	61.85	28.48	23.20	19.88
	5-shot	67.86	27.94	21.12	18.82

Context	en-fr	en-es
0-shot	20.92	21.32
1-shot		
Random	14.84	15.81
BM25	21.42	22.80
3-shot		
Random	16.78	18.88
BM25	22.58	24.66
5-shot		
Random	17.69	20.23
BM25	23.15	25.58

METHOD

Large language model



Absolute: (margin(a, b) = a)

Relative: (margin(a, b) = $a-b$)

Ratio (margin(a, b) = a / b)

At the sentence level: based on boundary similarity calculation method, considering the semantic centroid shift, selecting the most appropriate sentence by finding the average of k similar sentences to the similar sentence.

METHOD

Translate the following English sentence to Spanish: Some of the rainfall was accompanied by thunderstorms and frequent lightning.

Wherein, the English word "rainfall" translates Spanish word "precipitaciones" is 0.8; the English word "lightning" and Spanish word "relámpagos";

The translation to Spanish is: Parte de las precipitaciones contaron, además, con tormentas eléctricas y numerosos relámpagos.

...

Translate the following English sentence to Spanish: No one was inside the apartment.

Wherein, the similarity of English word "apartment" and Spanish word "apartamento" is 0.9;

The translation to Spanish is:

While designing sentence-level prompts based on sentence-level similarity, we also considered incorporating word-level prompts.

$$\text{Score}_w^{st}(\mathbf{w}_i^{\mathbf{x}}.\mathbf{w}_j^{\hat{\mathbf{y}}}) = \cos(\mathbf{w}_i^{\mathbf{x}}.\mathbf{w}_j^{\hat{\mathbf{y}}}) > \alpha \qquad \text{Score}_w^{ts}(\mathbf{w}_j^{\hat{\mathbf{y}}}.\mathbf{w}_m^{\hat{\mathbf{x}}}) = \cos(\mathbf{w}_j^{\hat{\mathbf{y}}}.\mathbf{w}_m^{\hat{\mathbf{x}}}) > \beta$$

Experiments

Methods	en-fr		fr-en		en-es		es-en		en-pt		pt-en	
	O	F	O	F	O	F	O	F	O	F	O	F
0-shot	20.9	32.9	17.0	34.7	21.3	21.7	24.2	36.9	13.4	27.0	19.7	38.6
1-shot												
Random	14.8	25.0	15.7	31.6	15.8	15.0	16.6	20.1	12.2	25.7	16.2	24.1
BM25	21.4	32.5	21.3	38.4	22.8	20.0	21.4	24.4	17.2	31.3	18.3	31.4
Fuzzy	21.6	38.7	21.8	43.2	23.0	21.8	22.3	28.1	17.6	36.1	21.6	35.2
Absolute	22.4	40.7	22.3	45.2	23.5	20.7	26.4	31.0	19.2	37.4	23.9	41.5
Relative	22.8	41.1	21.9	**45.8**	**23.7**	21.1	**27.1**	31.3	19.9	37.9	**24.5**	41.8
Ratio	**23.2**	**41.7**	**22.5**	45.6	23.1	**21.5**	26.8	**31.8**	**20.1**	**38.1**	24.2	**42.1**
3-shot												
Random	16.8	35.4	15.2	33.0	18.9	21.2	18.6	23.0	16.6	35.4	17.2	27.8
BM25	22.6	41.7	22.4	38.2	24.7	23.8	23.9	25.4	19.5	38.7	21.3	33.7
Fuzzy	22.3	43.7	23.3	42.6	23.7	24.8	22.9	29.0	19.4	39.5	21.3	39.5
Absolute	23.3	44.3	25.7	49.6	27.0	25.7	30.2	39.3	22.1	45.1	26.2	50.9
Relative	23.9	**45.3**	25.9	**51.1**	**27.6**	26.2	30.9	39.8	23.1	45.6	26.5	**51.8**
Ratio	**24.1**	45.2	**26.0**	50.8	27.3	**26.5**	**31.1**	**40.2**	**23.5**	**46.1**	**26.8**	51.3
5-shot												
Random	17.7	39.4	16.4	34.9	20.2	23.4	19.1	25.0	17.2	39.6	17.9	27.8
BM25	23.2	42.5	21.8	37.8	25.6	24.2	22.7	27.0	20.3	40.0	21.8	34.7
Fuzzy	22.3	43.4	22.8	42.9	22.5	25.2	24.1	28.9	18.7	40.4	22.5	39.4
Absolute	22.9	48.1	27.1	52.4	27.7	27.8	31.0	40.7	22.1	46.7	27.6	52.8
Relative	**23.3**	48.9	**27.7**	**53.0**	**28.3**	**28.6**	**31.9**	41.3	22.5	**47.4**	28.1	53.3
Ratio	23.2	**49.2**	27.5	52.8	28.1	28.4	31.5	**41.4**	**22.6**	47.3	**28.5**	**53.6**

Experiments

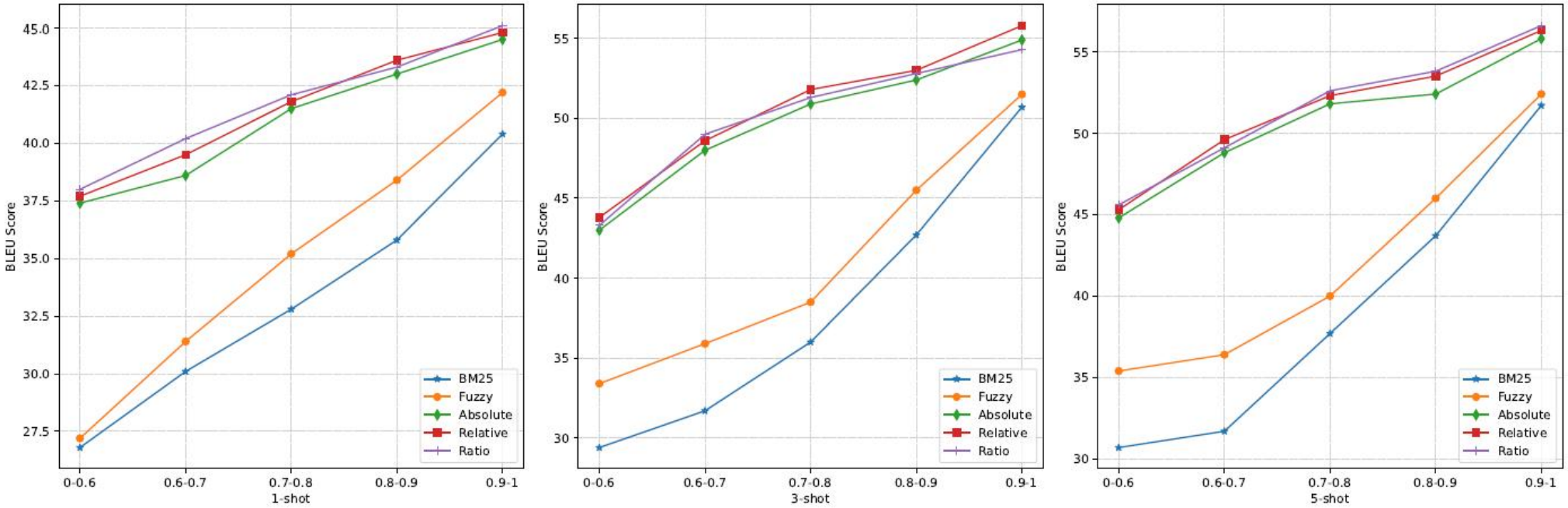
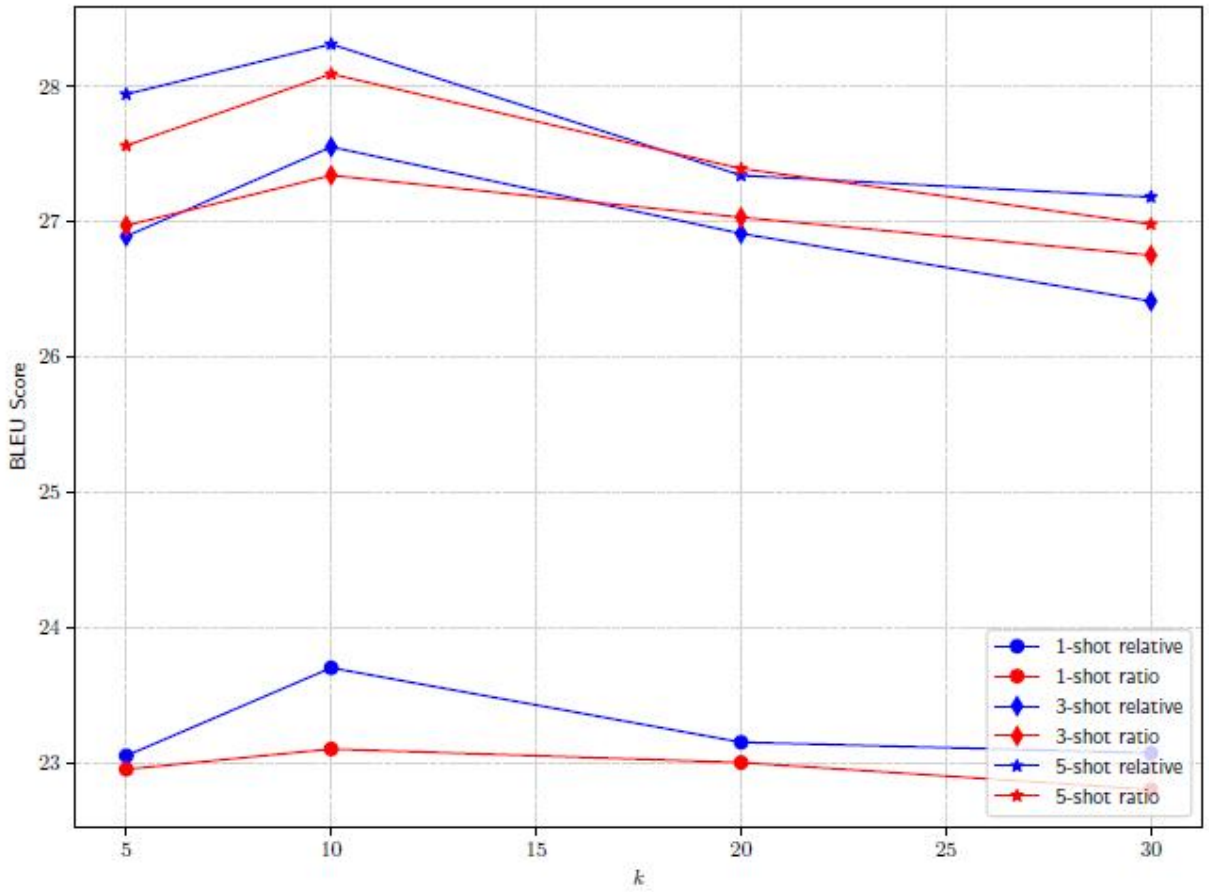


Figure 1: The robustness of different methods on pt-en language pair of Flores-200 test set.

We tested the robustness of large language models in performing translation tasks on sample pairs under different similarity conditions. One clear conclusion is that under relatively high similarity, there is no significant difference in the translation abilities of different methods. However, when the similarity is low, our method demonstrates greater stability.

Experiments



We tested the impact of selecting different k samples for similarity calculation, and concluded that introducing a large number of sentences will lead to the impact of irrelevant information and reduce the performance of translation.

Experiments

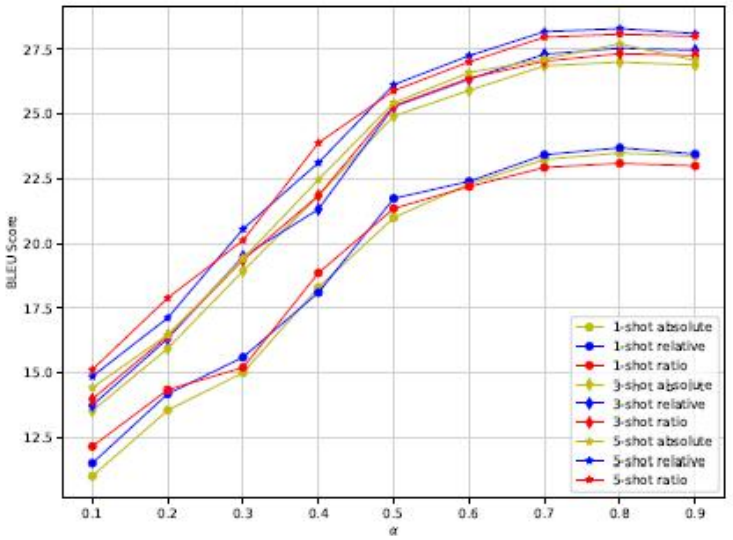


Figure 3: The BLEU score of our method on the OPUS test set for en-es with different α.

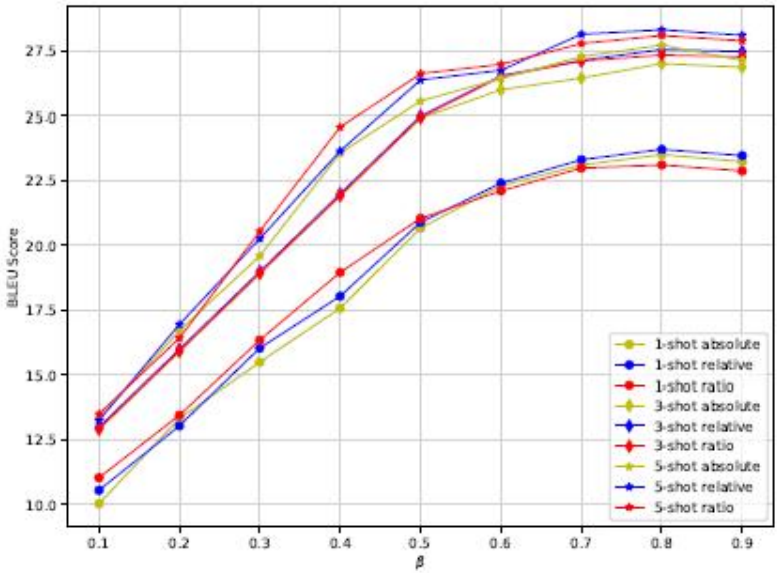


Figure 4: The BLEU score of our method on the OPUS test set for en-es with different β.

We further tested the impact of different selections of word—level information with different similarities on translation. It can be clearly concluded that noisy data significantly limits the translation of large models.

天津大學
Tianjin University

THANKS!