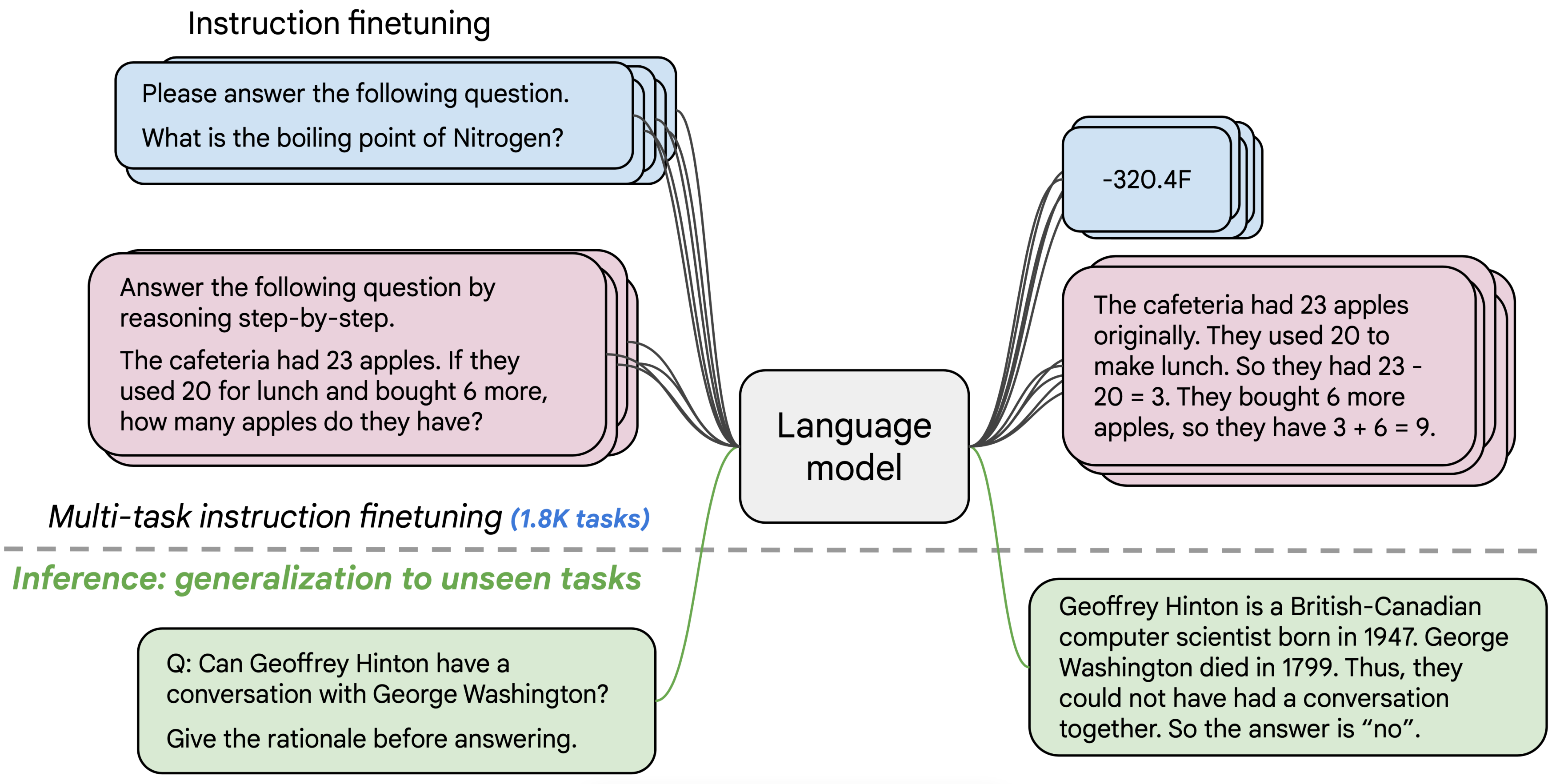
Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors

Kai Zhang, Bernal Jiménez Gutiérrez, Yu Su

The Ohio State University



Chung et al., Scaling Instruction-Finetuned Language Models, 2022

T0

BigScience

"Finally, in explaining the success of prompts, the leading hypothesis is that models learn to understand the prompts as task instructions which help them generalize to held-out tasks"

Sanh et. al. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization

FLAN

Google

"We show that instruction tuning—finetuning language models on a collection of datasets described via instructions—substantially improves zero-shot performance on unseen tasks."

Wei et. al. 2022. Finetuned Language Models Are Zero-Shot Learners.

InstructGPT

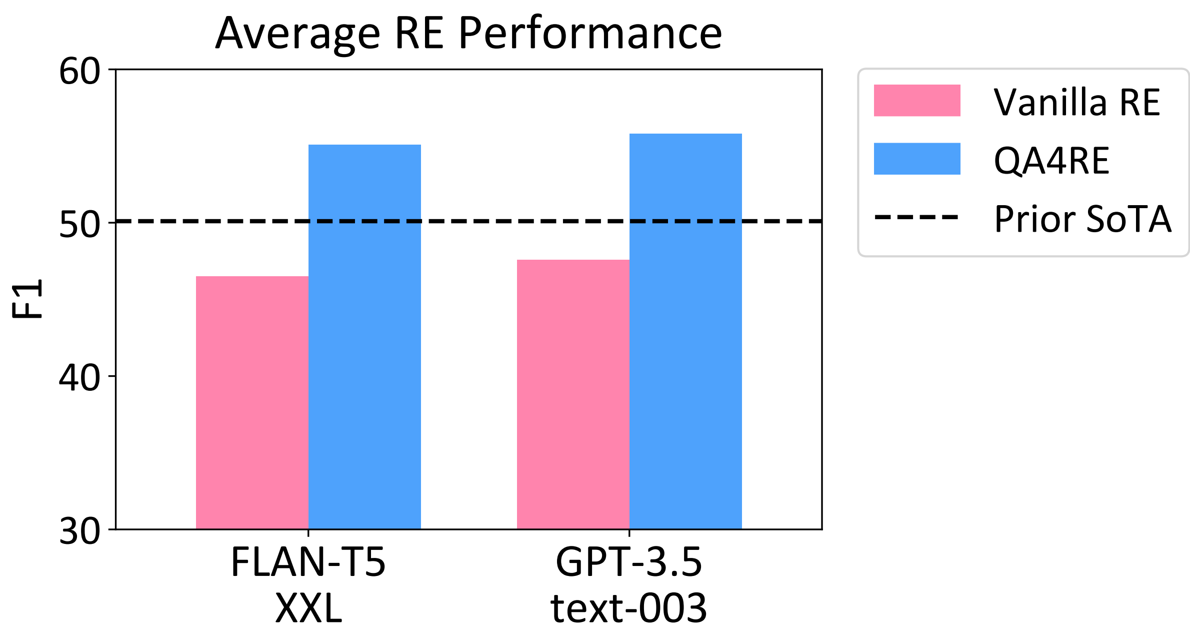
OpenAI

"A consistent finding across studies is that fine-tuning LMs on a range of NLP tasks, with instructions, improves their downstream performance on held-out tasks, both in the zero-shot and few-shot settings."

Ouyang et. al. 2022. Training language models to follow instructions with human feedback

Motivating Observation

- Instruction-tuned LLMs underperform small LMs methods on relation extraction (RE)



We regard language models with <1B parameters as small.

Reason: Lack of RE-like Instruction Tuning?

	#Tasks	%RE	%QA
T0 Collection (Sanh et al., 2022)	62	0	27.4
FLAN Collection (Wei et al., 2022)	62	0	21
MetaICL Collection (Min et al., 2022)	142	0	28.9
Natural Instructions v2 (Wang et al., 2022)	1731	<0.5	>12

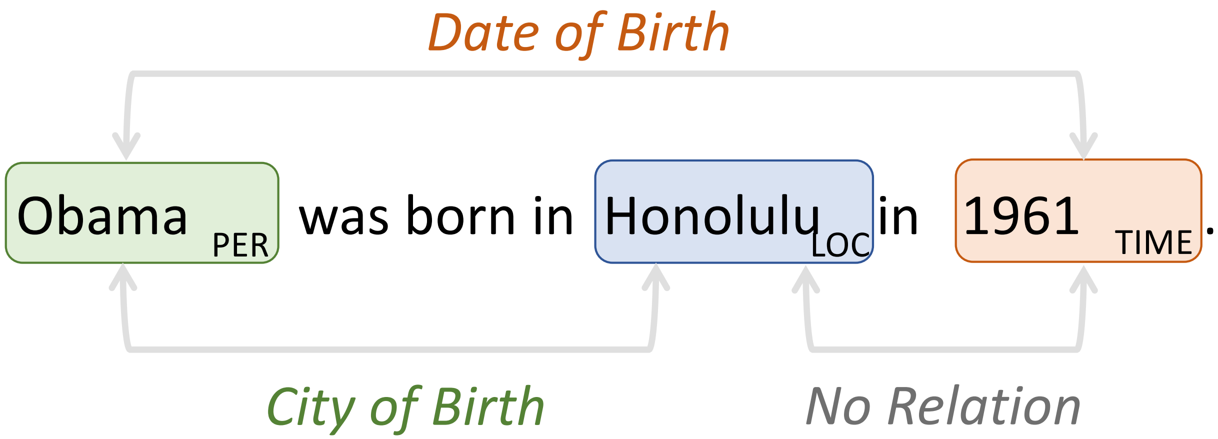
Table 1: Prevalent instruction tuning collections and proportion of RE and QA tasks in each.

Research Question

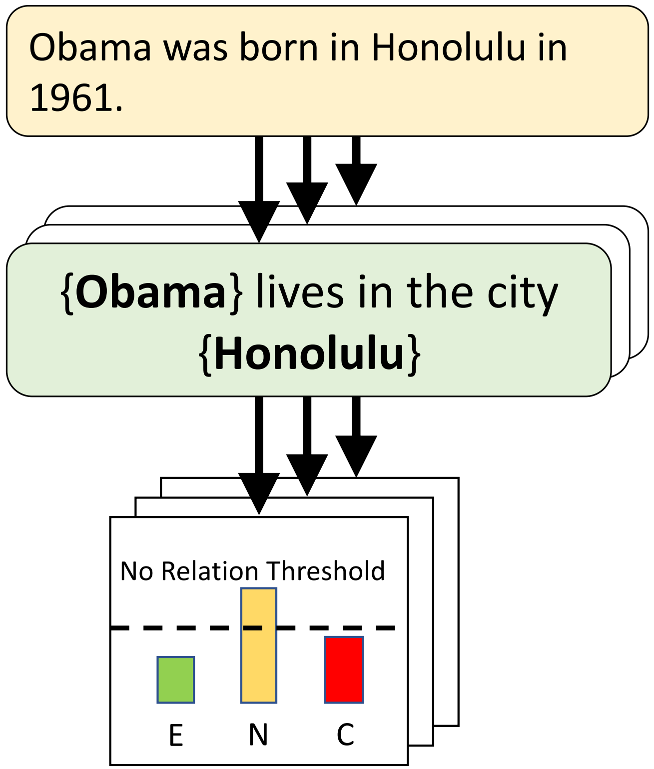
- How LLMs will perform with unpopular (RE) and popular (QA) instructions & formats?

- -> Are unpopular (RE) and popular (QA) tasks equally benefit from instruction tuning?

Relation Extraction



SoTA zero-shot RE: NLI



Vanilla RE

Given a sentence, and two entities within the sentence, classify the relationship between the two entities based on the provided sentence. All possible relationships are listed below:
- per:city_of_birth
- per:city_of_death
- per:cities_of_residence
- no_relation

Sentence: **Wearing jeans and a white blouse, Amanda Knox of Seattle is being cross-examined by prosecutors.**
Entity 1 : **Amanda Knox**
Entity 2 : **Seattle**
Relationship: **per:city_of_birth** ✗

QA4RE

Determine which option can be inferred from the given sentence.

Sentence: **Wearing jeans and a white blouse, Amanda Knox of Seattle is being cross-examined by prosecutors.**

Options:
A. Amanda Knox was born in the city Seattle
B. Amanda Knox died in the city Seattle
C. Amanda Knox lives in the city Seattle
D. Amanda Knox has no known relations to Seattle

Which option can be inferred from the given sentence?
Option: **C.** ✓

Gutiérrez et al., Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again, 2022

Methods		TACRED			RETACRED			TACREV			SemEval			Avg.
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
Baselines														
NLI_{BART}		42.6	65.0	51.4	59.5	34.9	44.0	44.0	74.6	55.3	21.6	23.7	22.6	43.3
$\text{NLI}_{\text{RoBERTa}}$		37.1	76.9	50.1	52.3	67.0	58.7	37.1	83.6	51.4	17.6	20.9	19.1	44.8
$\text{NLI}_{\text{DeBERTa}}$		42.9	76.9	55.1	71.7	58.3	64.3	43.3	84.6	57.2	22.0	25.7	23.7	50.1
$\text{SuRE}_{\text{BART}}$		13.1	45.7	20.4	17.9	34.6	23.6	14.1	52.3	22.2	0.0	0.0	0.0	16.5
$\text{SuRE}_{\text{PEGASUS}}$		13.8	51.7	21.8	16.6	34.6	22.4	13.5	54.1	21.6	0.0	0.0	0.0	16.4
GPT-3.5 Series														
ChatGPT	Vanilla	32.1	74.8	44.9	45.4	61.3	52.1	30.3	79.6	43.9	18.2	20.8	19.4	40.1
code-002	Vanilla	27.2	70.1	39.2	42.7	70.4	53.1	27.5	77.7	40.6	27.2	25.6	26.4	39.8
text-002	Vanilla	31.2	73.1	43.7	44.1	76.3	55.9	30.2	76.8	43.3	31.4	28.8	30.1	43.2
text-003	Vanilla	36.9	68.8	48.1	49.7	62.2	55.3	38.2	76.8	51.0	33.2	39.3	36.0	47.6
FLAN-T5 Series														
XLarge	Vanilla	51.6	49.1	50.3	54.3	40.3	46.3	56.0	59.1	57.5	35.6	29.8	32.4	46.6
XXLarge	Vanilla	52.1	47.9	49.9	56.6	54.0	55.2	52.6	50.9	51.7	29.6	28.8	29.2	46.5

Table 1: Experimental results on four RE datasets (%). We omit the 'davinci' within the names of GPT-3.5 Series LLMs and ChatGPT refers to gpt-3.5-turbo-0301. We mark the best results in **bold**, the second-best underlined, and F1 improvement of our QA4RE over vanilla RE in green.

Methods		TACRED P	TACRED R	TACRED F1	RETACRED P	RETACRED R	RETACRED F1	TACREV P	TACREV R	TACREV F1	SemEval P	SemEval R	SemEval F1	Avg. F1
Baselines														
NLI_{BART}		42.6	65.0	51.4	59.5	34.9	44.0	44.0	74.6	55.3	21.6	23.7	22.6	43.3
$NLI_{RoBERTa}$		37.1	76.9	50.1	52.3	67.0	58.7	37.1	83.6	51.4	17.6	20.9	19.1	44.8
$NLI_{DeBERTa}$		42.9	76.9	<u>55.1</u>	71.7	58.3	64.3	43.3	84.6	57.2	22.0	25.7	23.7	50.1
$SuRE_{BART}$		13.1	45.7	20.4	17.9	34.6	23.6	14.1	52.3	22.2	0.0	0.0	0.0	16.5
$SuRE_{PEGASUS}$		13.8	51.7	21.8	16.6	34.6	22.4	13.5	54.1	21.6	0.0	0.0	0.0	16.4
GPT-3.5 Series														
ChatGPT	Vanilla	32.1	74.8	44.9	45.4	61.3	52.1	30.3	79.6	43.9	18.2	20.8	19.4	40.1
	QA4RE	32.8	68.0	44.2 (−0.7)	48.3	76.8	59.3 (+7.2)	34.7	79.1	48.2 (+4.3)	29.9	35.2	32.3 (+12.9)	46.0 (+5.9)
code-002	Vanilla	27.2	70.1	39.2	42.7	70.4	53.1	27.5	77.7	40.6	27.2	25.6	26.4	39.8
	QA4RE	37.7	65.4	47.8 (+8.6)	48.0	74.0	58.2 (+5.1)	31.7	65.5	42.7 (+2.1)	25.2	29.2	27.0 (+0.6)	43.9 (+4.1)
text-002	Vanilla	31.2	73.1	43.7	44.1	76.3	55.9	30.2	76.8	43.3	31.4	28.8	30.1	43.2
	QA4RE	35.6	68.4	46.8 (+3.1)	46.4	72.4	56.5 (+0.6)	35.7	76.8	48.8 (+5.4)	29.4	34.3	31.6 (+1.5)	45.9 (+2.7)
text-003	Vanilla	36.9	68.8	48.1	49.7	62.2	55.3	38.2	76.8	51.0	33.2	39.3	36.0	47.6
	QA4RE	47.7	78.6	**59.4** (+11.3)	56.2	67.2	61.2 (+5.9)	46.0	83.6	**59.4** (+8.4)	41.7	45.0	<u>43.3</u> (+7.3)	**55.8** (+8.2)
FLAN-T5 Series														
XLarge	Vanilla	51.6	49.1	50.3	54.3	40.3	46.3	56.0	59.1	57.5	35.6	29.8	32.4	46.6
XXLarge	Vanilla	52.1	47.9	49.9	56.6	54.0	55.2	52.6	50.9	51.7	29.6	28.8	29.2	46.5

Table 1: Experimental results on four RE datasets (%). We omit the 'davinci' within the names of GPT-3.5 Series LLMs and ChatGPT refers to gpt-3.5-turbo-0301. We mark the best results in **bold**, the second-best <u>underlined</u>, and F1 improvement of our QA4RE over vanilla RE in green.

Methods		TACRED			RETACRED			TACREV			SemEval			Avg.
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
Baselines														
NLI_{BART}		42.6	65.0	51.4	59.5	34.9	44.0	44.0	74.6	55.3	21.6	23.7	22.6	43.3
$NLI_{RoBERTa}$		37.1	76.9	50.1	52.3	67.0	58.7	37.1	83.6	51.4	17.6	20.9	19.1	44.8
$NLI_{DeBERTa}$		42.9	76.9	55.1	71.7	58.3	64.3	43.3	84.6	57.2	22.0	25.7	23.7	50.1
$SuRE_{BART}$		13.1	45.7	20.4	17.9	34.6	23.6	14.1	52.3	22.2	0.0	0.0	0.0	16.5
$SuRE_{PEGASUS}$		13.8	51.7	21.8	16.6	34.6	22.4	13.5	54.1	21.6	0.0	0.0	0.0	16.4
GPT-3.5 Series														
ChatGPT	Vanilla	32.1	74.8	44.9	45.4	61.3	52.1	30.3	79.6	43.9	18.2	20.8	19.4	40.1
	QA4RE	32.8	68.0	44.2 (−0.7)	48.3	76.8	59.3 (+7.2)	34.7	79.1	48.2 (+4.3)	29.9	35.2	32.3 (+12.9)	46.0 (+5.9)
code-002	Vanilla	27.2	70.1	39.2	42.7	70.4	53.1	27.5	77.7	40.6	27.2	25.6	26.4	39.8
	QA4RE	37.7	65.4	47.8 (+8.6)	48.0	74.0	58.2 (+5.1)	31.7	65.5	42.7 (+2.1)	25.2	29.2	27.0 (+0.6)	43.9 (+4.1)
text-002	Vanilla	31.2	73.1	43.7	44.1	76.3	55.9	30.2	76.8	43.3	31.4	28.8	30.1	43.2
	QA4RE	35.6	68.4	46.8 (+3.1)	46.4	72.4	56.5 (+0.6)	35.7	76.8	48.8 (+5.4)	29.4	34.3	31.6 (+1.5)	45.9 (+2.7)
text-003	Vanilla	36.9	68.8	48.1	49.7	62.2	55.3	38.2	76.8	51.0	33.2	39.3	36.0	47.6
	QA4RE	47.7	78.6	**59.4** (+11.3)	56.2	67.2	61.2 (+5.9)	46.0	83.6	**59.4** (+8.4)	41.7	45.0	43.3 (+7.3)	**55.8** (+8.2)
FLAN-T5 Series														
XLarge	Vanilla	51.6	49.1	50.3	54.3	40.3	46.3	56.0	59.1	57.5	35.6	29.8	32.4	46.6
	QA4RE	40.0	78.2	53.0 (+2.7)	57.1	79.7	66.5 (+20.2)	40.7	85.9	55.3 (−2.2)	45.1	40.1	42.5 (+10.1)	54.3 (+7.7)
XXLarge	Vanilla	52.1	47.9	49.9	56.6	54.0	55.2	52.6	50.9	51.7	29.6	28.8	29.2	46.5
	QA4RE	40.6	82.9	54.5 (+4.6)	56.6	82.9	**67.3** (+12.1)	39.6	86.4	54.3 (+2.6)	41.0	47.8	**44.1** (+14.9)	55.1 (+8.6)

Table 1: Experimental results on four RE datasets (%). We omit the 'davinci' within the names of GPT-3.5 Series LLMs and ChatGPT refers to gpt-3.5-turbo-0301. We mark the best results in **bold**, the second-best underlined, and F1 improvement of our QA4RE over vanilla RE in green.

Methods	TACRED			RETACRED			TACREV			SemEval			Avg.
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1
Baselines													
NLI_{BART}	42.6	65.0	51.4	59.5	34.9	44.0	44.0	74.6	55.3	21.6	23.7	22.6	43.3

Transforming the unfamiliar tasks (RE) of instruction-tuned LLMs to their familiar tasks (QA) could bring significant performance gains.

Table 1: Experimental results on four RE datasets (%). We omit the 'davinci' within the names of GPT-3.5 Series LLMs and ChatGPT refers to gpt-3.5-turbo-0301. We mark the best results in **bold**, the second-best underlined, and F1 improvement of our QA4RE over vanilla RE in green.

Only work on Large LMs?

LMs	Model Size	Avg. F1 Vanilla	QA4RE	Δ
GPT-3.5 Series				
text-001	175B	22.3	14.9	−7.4
code-002	175B	39.8	43.9	+4.1
text-002	175B	43.2	45.9	+2.7
text-003	175B	47.6	55.8	+8.2
FLAN-T5 Series				
Small	80M	19.5	25.0	+5.6
Base	250M	22.3	26.4	+4.2
Large	780M	34.8	41.8	+7.0
XLarge	3B	46.6	54.3	+7.7
XXLarge	11B	46.5	55.1	+8.6

Table 7: Effectiveness of QA4RE on both the GPT-3.5 series and FLAN-T5 with different sizes. The results are averaged over four RE datasets.

1. Effectively transferable from 80M (FLAN-T5 Small) to 175B (text-davinvi-003).
2. In GPT-3.5 series: the more recent model, the more performance gains from QA4RE.
3. In FLAN-T5: larger models, more performance gains from QA4RE.

Are Relation Templates All LLMs Need?

Vanilla + Template RE

Given a sentence, and two entities within the sentence, classify the relationship between the two entities based on the provided sentence. All possible relationships are listed below with explanations:
- per:city_of_birth: Entity 1 was born in the city Entity 2
- per:city_of_death: Entity 1 died in the city Entity 2
- per:cities_of_residence: Entity 1 lives in the city Entity 2
- no_relation: Entity 1 has no known relations to Entity 2

Sentence: **Wearing jeans and a white blouse, Amanda Knox of Seattle is being cross-examined by prosecutors.**
Entity 1 : **Amanda Knox**
Entity 2 : **Seattle**
Relationship: **per:city_of_birth**

Methods		P	R	F1	ΔF1
code-002	Vanilla	27.2	70.1	39.2	-
	Vanilla + TEMP	27.5	71.8	39.7	+0.5
	QA4RE	37.7	65.4	47.8	+8.6
text-002	Vanilla	31.2	73.1	43.7	-
	Vanilla + TEMP	26.8	77.8	39.8	−3.9
	QA4RE	35.6	68.4	46.8	+3.1
text-003	Vanilla	36.9	68.8	48.1	-
	Vanilla + TEMP	36.9	76.5	49.8	+1.7
	QA4RE	**47.7**	**78.6**	**59.4**	+11.3

Table 5: Evaluation on TACRED regarding whether incorporating relation explanations based on the same templates into vanilla RE bridges its gap to QA4RE (%).

How Strong the QA4RE is? - Robustness

Methods		TEMP1	TEMP2	TEMP3	TEMP4
$\mathrm{NLI}_{\mathrm{BART}}$		51.4	49.7	4.4	42.0
$\mathrm{NLI}_{\mathrm{RoBERTa}}$		50.1	47.1	19.6	35.8
$\mathrm{NLI}_{\mathrm{DeBERTa}}$		55.0	49.4	17.1	36.6
$\mathrm{SuRE}_{\mathrm{BART}}$		19.9	20.4	2.1	10.1
$\mathrm{SuRE}_{\mathrm{PEGASUS}}$		20.5	21.8	6.2	19.3
text-003	Vanilla	48.1			
	QA4RE	**56.6**	**59.4**	**48.7**	**50.1**

Table 2: F1 score on TACRED with four templates (%) The best result using each template is marked in bold. text-003 refers to text-davinci-003.

org:top_members/employees

1. Concrete Examples: *{E_h} is a chairman/ president/director of {E_t}*
2. Semantic Relationship: *{E_h} is a high level member of {E_t}*
3. Straightforward: *The relation between {E_h} and {E_t} is top members or employees*
4. Word Translation: *{E_h} organization top members or employees {E_t}*

How Strong the QA4RE is? - Few-shot

Methods	K=0	K=4	K=8	K=16	K=32
Fine-Tuning	-	9.0	21.2	29.3	33.9
PTR	-	26.8	30.0	32.9	36.8
KnowPrompt	-	30.2	33.7	34.9	35.0
$\text{NLI}_{\text{DeBERTa}}$-TEMP1	55.0	**64.2**	**64.7**	**58.7**	**65.7**
$\text{NLI}_{\text{DeBERTa}}$-TEMP2	49.4	**51.2**	47.3	**50.5**	48.1
Vanilla	48.1	46.2		-	
QA4RE	59.4	**62.0**		-	

Table 4: Few-shot F1 on TACRED (%). All results are averaged over 3 different training subsets for each K. We use text-davinci-003 for vanilla RE and QA4RE. For the best-performing baseline (NLI) as well as vanilla RE and QA4RE, we mark the results in **bold** when they are improved over their zero-shot alternatives.

Vanilla RE - Few-Shot

[Instruction]

Sentence: Obama was born in Honolulu in 1961.
Entity 1: Obama
Entity 2: Honolulu
Relationship: per:city_of_birth

Sentence: **Wearing jeans ….**
Entity 1 : **Amanda Knox**
Entity 2 : **Seattle**
Relationship:

1. Vanilla RE do not benefit from few-shot demonstrations.
2. NLI baseline is still sensitive to templates in few-shot setting

Conclusion

- **1.** Reformulating tasks (RE) that are not well covered in the instruction datasets to popular tasks (QA) unlocks LLMs' abilities.

- **2.** QA4RE makes LLMs strong and robust zero-shot relation extractors.

Check our paper for more details 🤗 👉