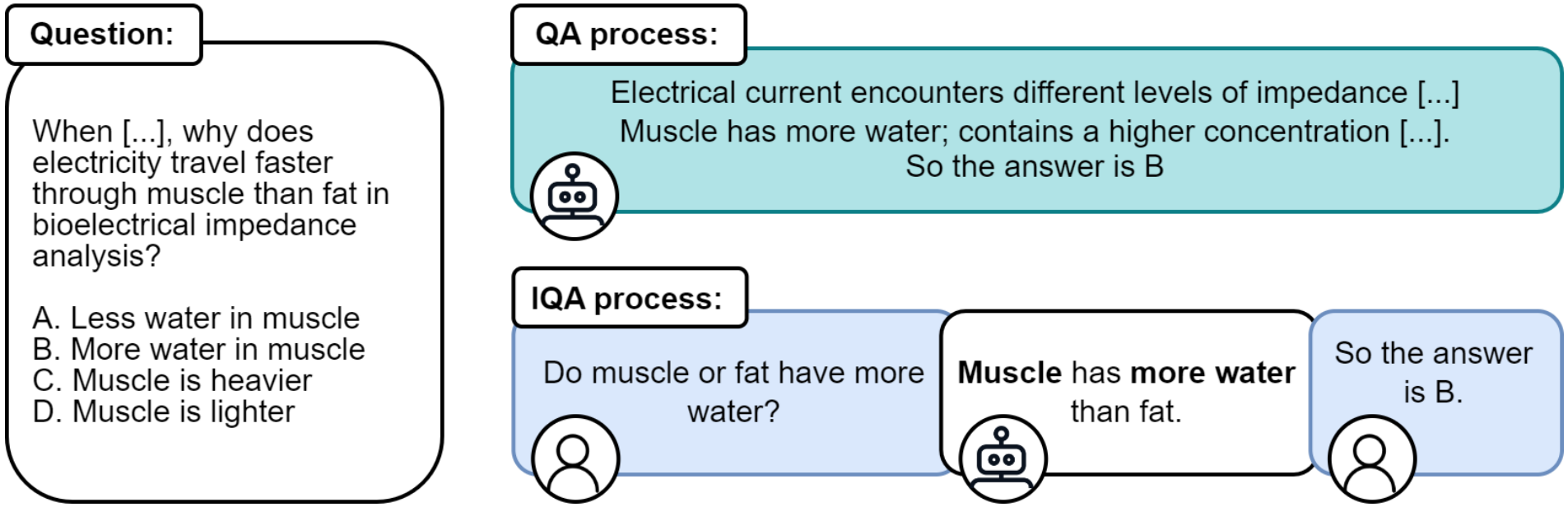
IQA-EVAL: Automatic Evaluation of Human-Model Interactive Question Answering

Ruosen Li[1], Ruochen Li[1], Barry Wang[2], Xinya Du[1]

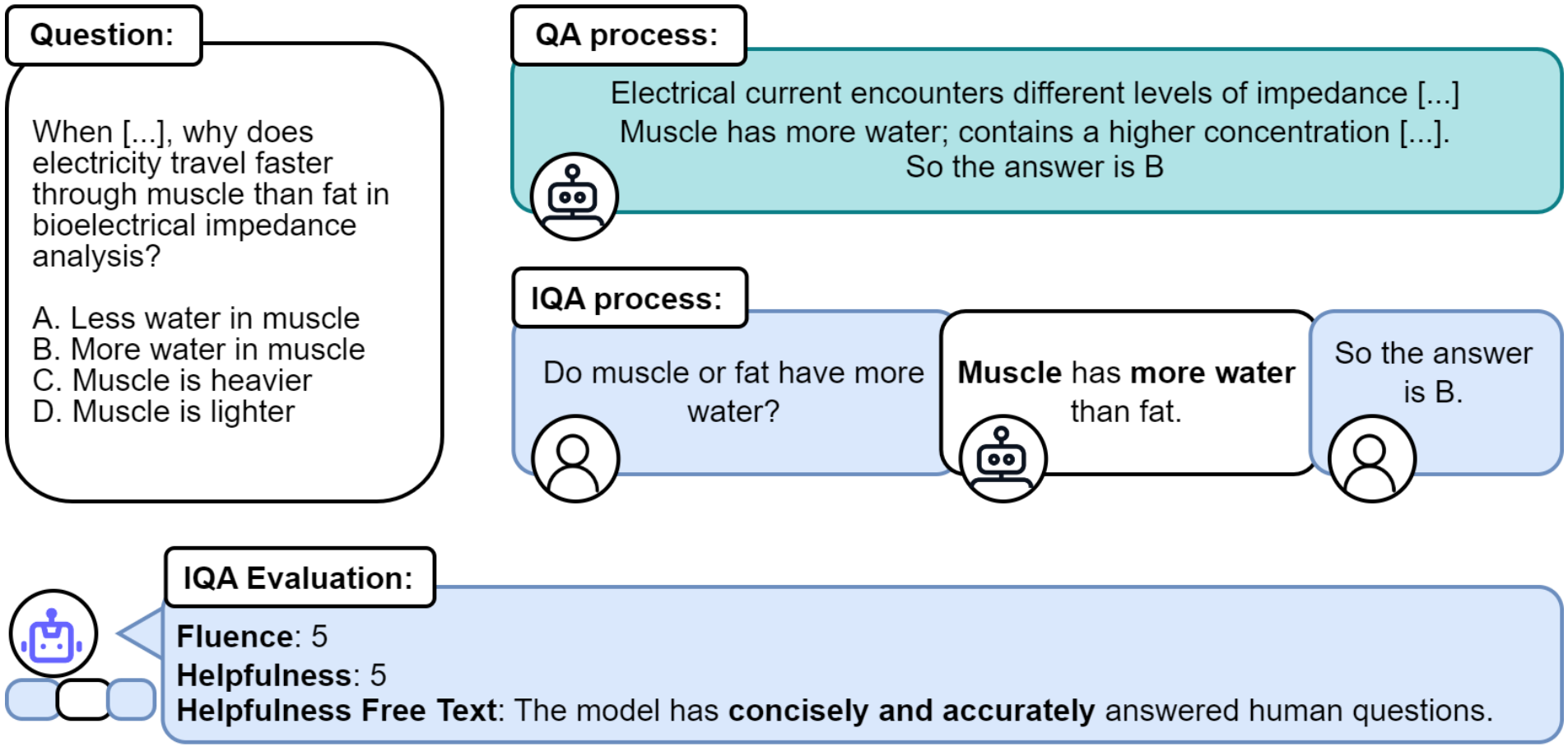
[1]Department of Computer Science, University of Texas at Dallas

[2]Department of Computer Science, Carnegie Mellon University

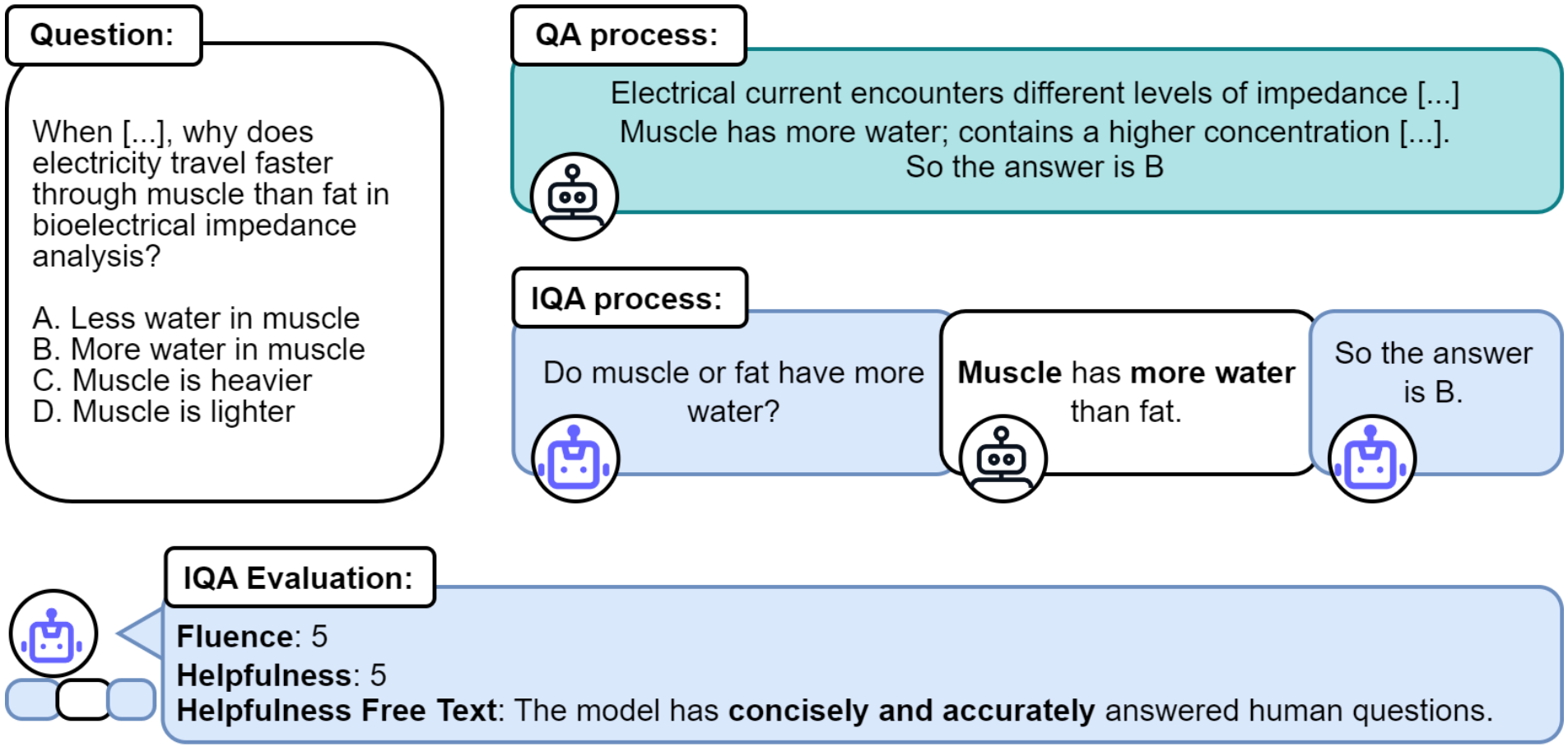
Interactive Question Answering



Interactive Question Answering Evaluation



Interactive Question Answering Evaluation



IQA-Eval Results

Table 1: IQA-EVAL evaluation results of IQA models (TDA: `TextDavinci`; TB: `TextBabbage`; DA: `Davinci`). **Bold numbers** indicate they are the most close to human results. The empty set symbol (Ø) indicates the number cannot be calculated due to the model's inability to follow instructions and produce a gradable answer.

	Helpfulness			**Fluency**			**# Queries**			**Accuracy**		
Evaluator	TDA	TB	DA	TDA	TB	DA	TDA	TB	DA	TDA	TB	DA
Human	4.60	3.84	3.52	4.35	3.84	3.22	1.78	2.57	2.66	69.00	52.00	48.00
IQA-EVAL-GPT4	3.67	2.30	2.10	4.77	**3.87**	3.03	**1.57**	2.27	**2.37**	0.87	0.83	0.67
IQA-EVAL-Claude	4.13	3.03	3.00	**4.47**	3.47	**3.23**	2.20	**2.67**	2.07	**0.67**	**0.53**	0.57
IQA-EVAL-GPT3.5	**4.30**	**3.87**	**3.93**	**4.47**	3.67	3.97	**1.57**	1.77	2.00	0.63	0.47	**0.53**

IQA-Eval Results

Table 1: IQA-EVAL evaluation results of IQA models (TDA: `TextDavinci`; TB: `TextBabbage`; DA: `Davinci`). **Bold numbers** indicate they are the most close to human results. The empty set symbol (Ø) indicates the number cannot be calculated due to the model's inability to follow instructions and produce a gradable answer.

	Helpfulness			**Fluency**			**# Queries**			**Accuracy**		
Evaluator	TDA	TB	DA	TDA	TB	DA	TDA	TB	DA	TDA	TB	DA
Human	4.60	3.84	3.52	4.35	3.84	3.22	1.78	2.57	2.66	69.00	52.00	48.00
IQA-EVAL-GPT4	3.67	2.30	2.10	4.77	**3.87**	3.03	**1.57**	2.27	**2.37**	0.87	0.83	0.67
IQA-EVAL-Claude	4.13	3.03	3.00									
IQA-EVAL-GPT3.5	**4.30**	**3.87**	**3.93**									

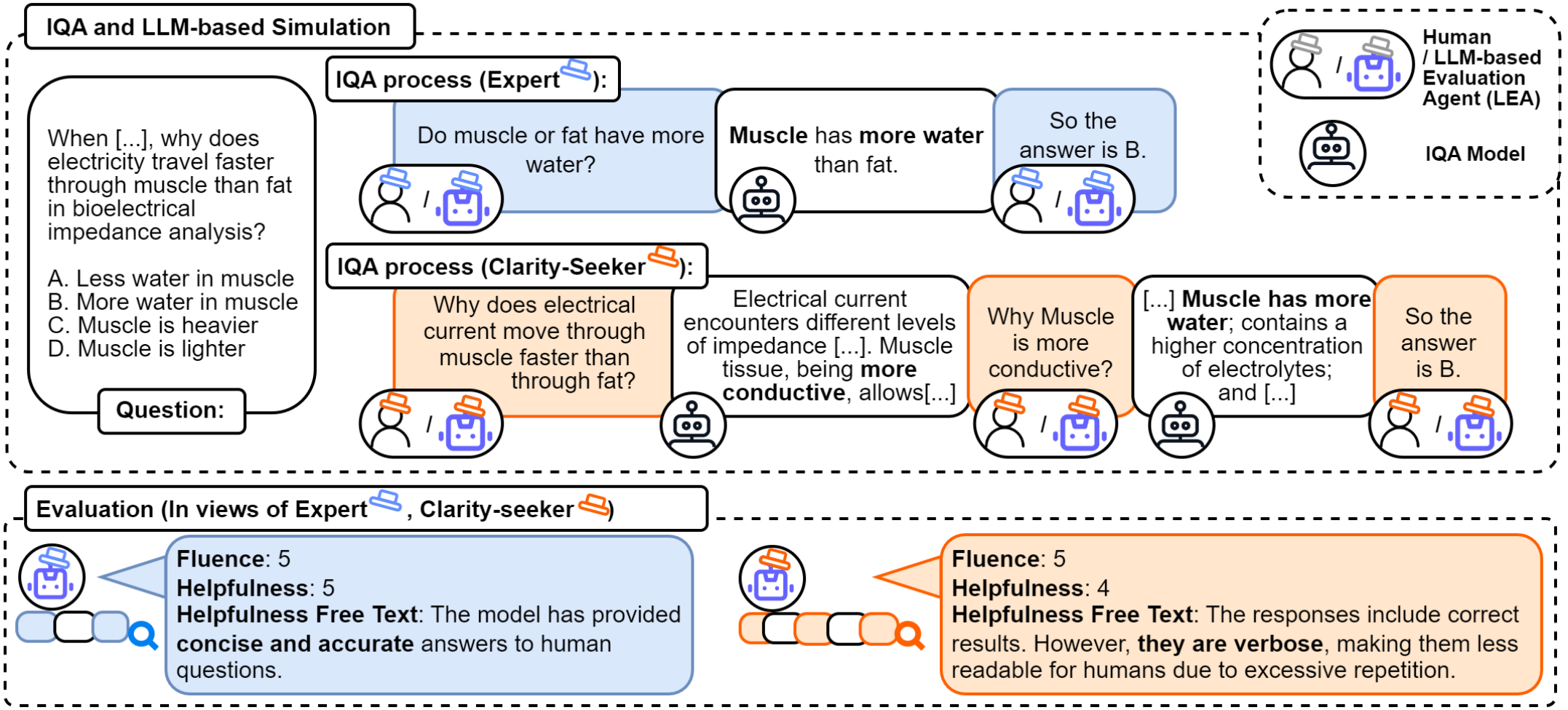
Table 2: Pearson Correlation (ρ) between IQA-EVAL evaluations and human judgments.

	Helpfulness	**Fluency**	**Overall**
IQA-EVAL-GPT4	**0.652**	**0.591**	**0.613**
IQA-EVAL-Claude	0.640	0.552	0.551
IQA-EVAL-GPT3.5	0.621	0.523	0.510

Personas

- **Expert**: knowledgeable; quickly learns new concepts and applies them in the reasoning process to answer questions.
- **Critical-Thinker**: people who prefer critical information rather than redundant or detailed responses.
- **Adaptability-Seeker**: people who prefer assistants can understand their questions even if they are not precise.
- **Clarity-Seeker**: people who prefer clear explanations from assistants.

Interactive Question Answering Evaluation (w/ Persona)



Effect of Assigning Persona

Table 3: IQA-EVAL evaluation results of IQA models (TDA: `TextDavinci`; TB: `TextBabbage`; DA: `Davinci`). LEAs, based on `GPT3.5`, are assigned specific personas when representing specific groups of workers.

Evaluator	# Queries			Accuracy		
	TDA	TB	TD	TDA	TB	TD
Human	1.78	2.57	2.66	0.69	0.52	0.48
IQA-EVAL	1.57	1.77	2.00	0.63	0.47	0.53
IQA-EVAL (Expert)	1.20	1.49	2.20	**0.73**	0.56	0.53
IQA-EVAL (Critical-Thinker)	1.55	1.80	1.99	0.68	0.54	0.55
IQA-EVAL (Adaptability-Seeker)	1.50	1.75	2.10	0.66	0.52	0.55
IQA-EVAL (Clarity-Seeker)	1.64	2.10	2.34	0.63	**0.57**	**0.57**

Effect of Assigning Persona

Table 4: IQA-EVAL evaluation results (helpfulness and fluency) of IQA models. Correlations are between the LEA evaluation in each row and human evaluations.

Evaluator	Helpfulness				Fluency				Overall
	TDA	TB	TD	ρ	TDA	TB	TD	ρ	ρ
Human	4.60	3.84	3.52	-	4.35	3.84	3.22	-	-
IQA-EVAL	4.30 (±0.06)	3.87 (±0.11)	3.93 (±0.13)	0.621	4.47 (±0.05)	3.67 (±0.08)	3.97 (±0.06)	0.523	0.510
IQA-EVAL (Expert)	4.17 (±0.08)	3.08 (±0.09)	3.12 (±0.11)	0.756	4.47 (±0.02)	3.84 (±0.04)	3.40 (±0.04)	0.787	0.670
IQA-EVAL (Critical-Thinker)	4.44 (±0.08)	4.02 (±0.13)	4.08 (±0.17)	0.711	4.64 (±0.06)	3.97 (±0.08)	4.10 (±0.08)	0.624	0.634
IQA-EVAL (Adaptability-Seeker)	4.24 (±0.05)	3.67 (±0.11)	3.75 (±0.11)	0.713	4.52 (±0.08)	3.84 (±0.07)	3.84 (±0.09)	0.637	0.650
IQA-EVAL (Clarity-Seeker)	4.45 (±0.07)	3.77 (±0.15)	3.80 (±0.12)	0.747	4.60 (±0.04)	3.85 (±0.04)	3.94 (±0.06)	0.676	0.690

Effect of Assigning Persona

Table 15: IQA-EVAL results under different persona distribution on the expert persona.

LEA models	Helpfulness				Fluency			
	TDA	TB	DA	ρ	TDA	TB	DA	ρ
Human	4.60	3.84	3.52		4.35	3.84	3.22	
IQA-EVAL (Expert)	4.17	3.08	3.12	0.756	4.47	3.84	3.40	0.787
IQA-EVAL (20% Expert)	4.31	3.26	3.44	0.708	4.62	4.09	3.65	0.741
IQA-EVAL (40% Expert)	4.21	3.14	3.23	0.751	4.49	3.88	3.44	0.779
IQA-EVAL (60% Expert)	4.11	3.01	3.00	0.725	4.43	3.77	3.34	0.734
IQA-EVAL (80% Expert)	4.02	2.90	2.79	0.680	4.30	3.56	3.12	0.703
Human (Pure Expert)	4.69	4.00	3.73		4.36	3.96	3.26	
IQA-EVAL (Pure Expert)	4.37	3.57	3.33	0.778	4.20	3.40	2.97	0.786

Benchmarking LLMs

Table 5: IQA-EVAL benchmarking results on HotpotQA and AmbigQA datasets.

IQA Models	HotpotQA				AmbigQA			
	Helpfulness ↑	Fluency↑	# Queries↓	Accuracy↑	Helpfulness	Fluency	# Queries	Accuracy
TextDavinci	4.72	4.87	1.22	0.45	-	-	-	-
TextBabbage	4.70	4.88	1.74	0.37	-	-	-	-
Davinci	4.27	4.52	1.68	0.32	-	-	-	-
GPT3.5	4.72	4.95	1.49	0.63	4.91	4.97	1.89	0.60
GPT4	4.78	4.96	1.12	0.66	4.89	4.95	1.06	0.72
Claude	4.82	4.99	1.26	0.58	4.89	4.94	1.36	0.62
Llama2	4.70	4.95	1.32	0.55	4.96	4.94	1.79	0.52
Zephyr	4.64	4.88	1.01	0.40	4.38	4.66	1.03	0.45

Thank you
for hearing!