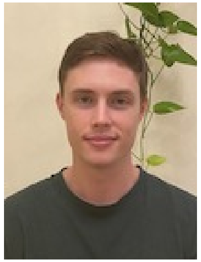
Retrieval-Augmented Debating (RAD)

Touché’25 Task 1

Marcel
Gohsen

Nailia
Mirzakhmedova

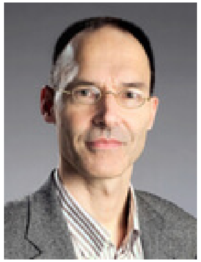


Harrisen
Scells

Mohammad
Aliannejadi

Maik
Fröbe

Johannes
Kiesel



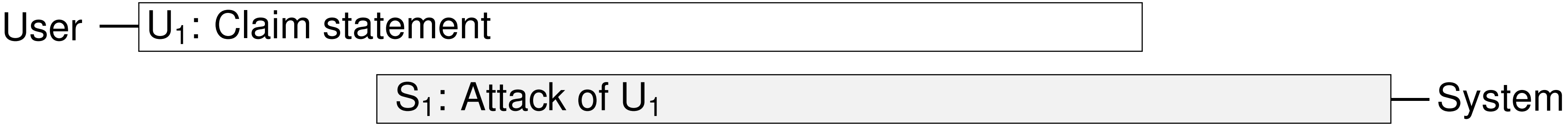
Benno
Stein

Retrieval-Augmented Debating (RAD)

User

U_1: Claim statement

Retrieval-Augmented Debating (RAD)



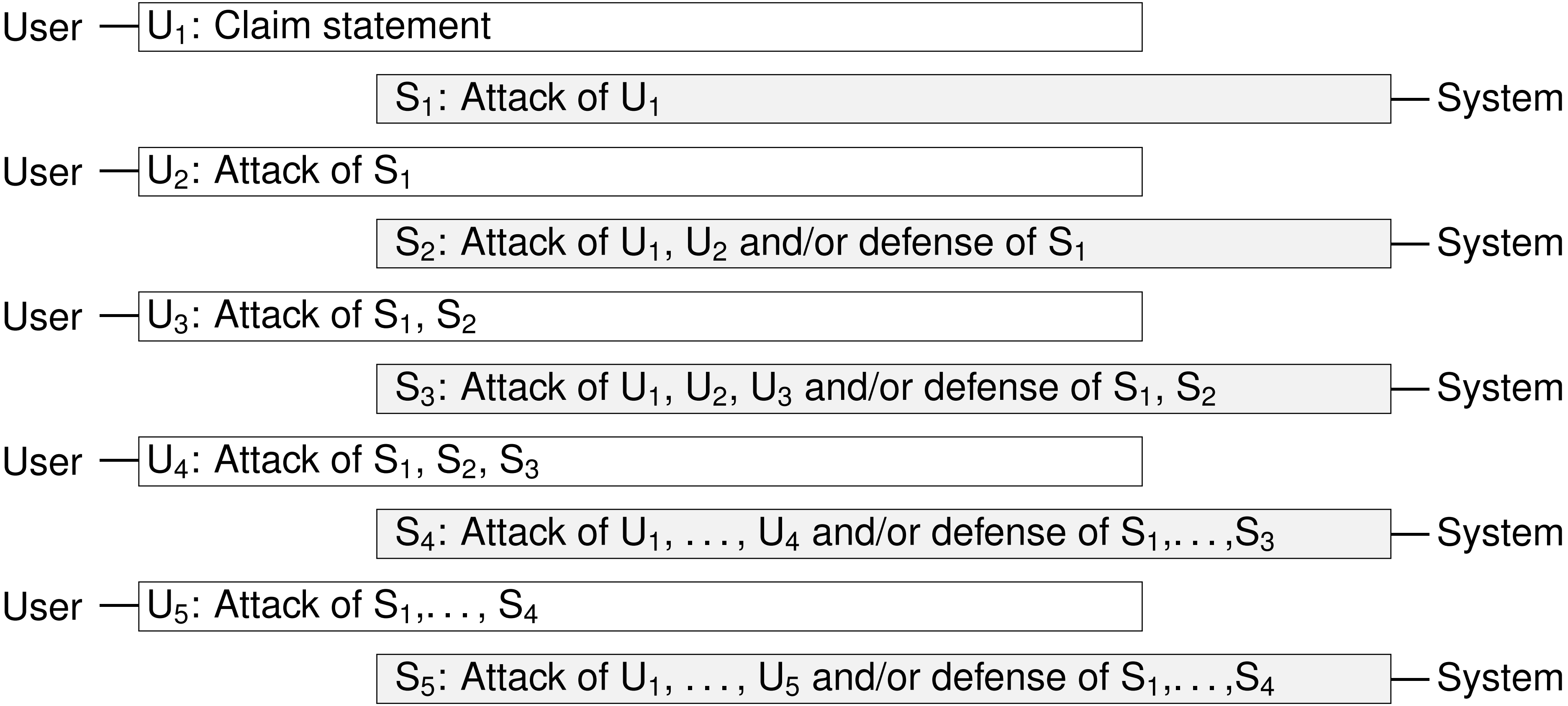
User

U_1: Claim statement

S_1: Attack of U_1

System

Retrieval-Augmented Debating (RAD)



User — U_1: Claim statement

S_1: Attack of U_1 — System

User — U_2: Attack of S_1

S_2: Attack of U_1, U_2 and/or defense of S_1 — System

User — U_3: Attack of S_1, S_2

S_3: Attack of U_1, U_2, U_3 and/or defense of S_1, S_2 — System

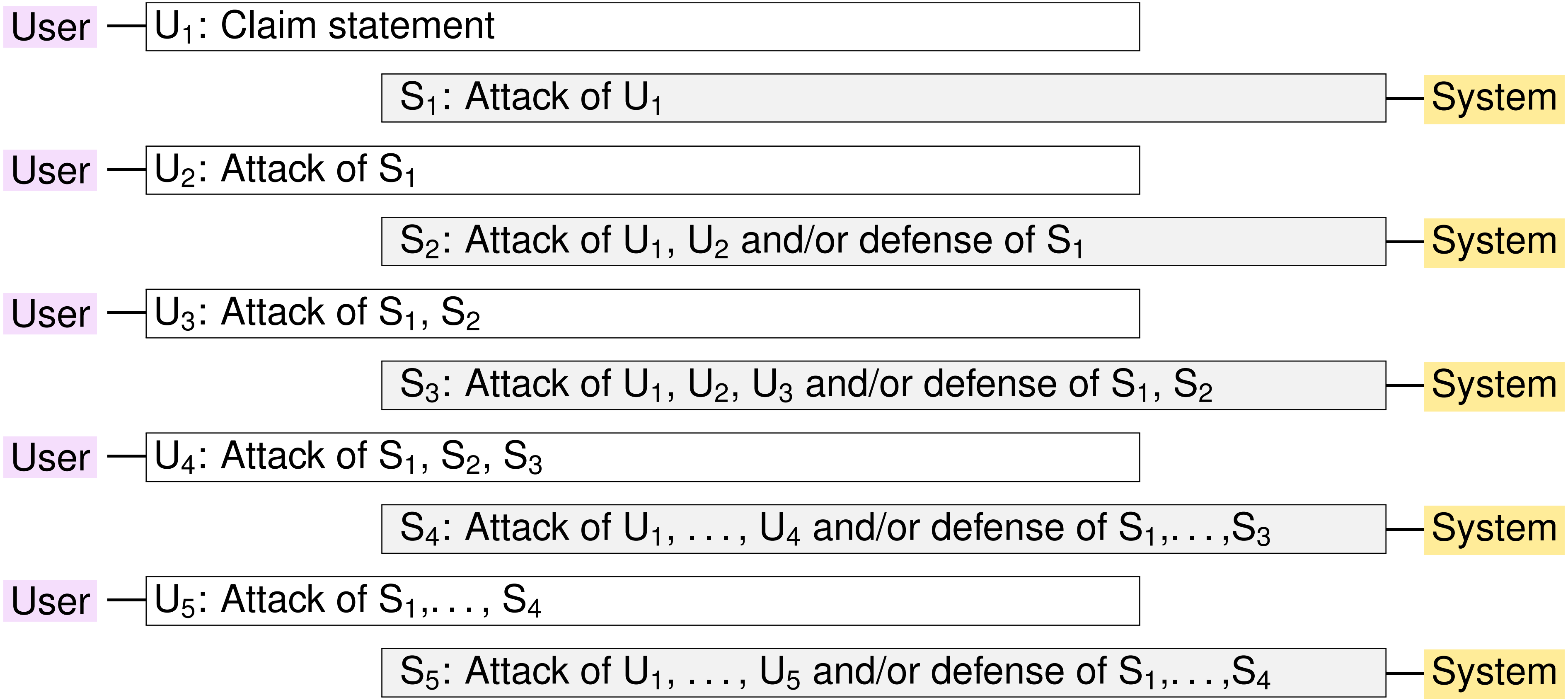
User — U_4: Attack of S_1, S_2, S_3

S_4: Attack of $U_1, \ldots, U_4$ and/or defense of $S_1, \ldots, S_3$ — System

User — U_5: Attack of $S_1, \ldots, S_4$

S_5: Attack of $U_1, \ldots, U_5$ and/or defense of $S_1, \ldots, S_4$ — System

Retrieval-Augmented Debating (RAD)



User — U_1: Claim statement

S_1: Attack of U_1 — System

User — U_2: Attack of S_1

S_2: Attack of U_1, U_2 and/or defense of S_1 — System

User — U_3: Attack of S_1, S_2

S_3: Attack of U_1, U_2, U_3 and/or defense of S_1, S_2 — System

User — U_4: Attack of S_1, S_2, S_3

S_4: Attack of U_1, . . . , U_4 and/or defense of S_1,. . . ,S_3 — System

User — U_5: Attack of S_1,. . . , S_4

S_5: Attack of U_1, . . . , U_5 and/or defense of S_1,. . . ,S_4 — System

🟪 User simulator 🟨 Participant system

Retrieval-Augmented Debating (RAD)

Task Description

Scenario: Assisting people in forming an opinion on controversial topics and training argumentation skills

Sub-Task 1: Develop debate systems that retrieve and respond with counterarguments and evidence in simulated debates.

Sub-Task 2: Provide metrics to assess quality criteria based on Grice’s maxims of cooperation.

Quantity: at least one at most one of each attack/defense arguments?

Quality: response grounded on retrieved arguments?

Relation: response coherent with conversation?

Manner: response clear and precise?

Retrieval-Augmented Debating (RAD)

Dataset

Arguments

- 300 000 arguments from ClaimRev[1]
- Pre-indexed in Elasticsearch

Argument: *Pineapple on pizza is an insult to the Italian origins of pizza.*

Supports: *Pineapple does not belong on pizza.*

Attacks: *Pineapple belongs on pizza.*

Claims and debates

- 100 claims from the Change My View subreddit[2]
- 100 simulated debates for claims with annotations
- Annotation: binary labels for quality criteria

[1] Skitalinskaya et al., Quality Assessment of Claims in Argumentation at Scale. EACL 2021.

[2] https://www.reddit.com/r/changemyview/

Retrieval-Augmented Debating (RAD)

Results: Sub-Task 1

Rank	Team	Run	Score	Quantity	Quality	Relation	Manner
1	DS@GT	gpt-4.1	**0.70**	**0.95**	0.17	0.82	**0.84**
2	DS@GT	gemini-2.5	0.65	0.94	0.26	0.74	0.67
	org	baseline	0.62	0.35	**1.00**	0.32	0.80
3	SINAI	run	0.54	0.70	0.02	0.86	0.59
4	DS@GT	gemini-2.5-flash	0.50	0.70	0.07	0.80	0.41
5	DS@GT	claude-opus-4	0.42	0.41	0.31	0.87	0.09
6	DS@GT	gpt-4o	0.42	0.20	0.02	0.86	0.58
7	DS@GT	claude-sonnet-4	0.38	0.35	0.05	**0.94**	0.17

Criteria: percentages of responses that fulfill given criteria.

Score: Avg. percentage of responses across all criteria.

Retrieval-Augmented Debating (RAD)

Results: Sub-Task 2

Rank	Team	Run	Score	Quantity			Quality			Relation			Manner		
			F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
	org	1-baseline	**0.67**	0.57	1.00	**0.73**	0.24	1.00	**0.38**	0.78	1.00	0.87	0.52	1.00	**0.68**
1	DS@GT	gemini-2.5-flash	0.64	0.59	0.86	0.70	0.18	0.66	0.29	0.81	0.99	0.89	0.52	0.99	**0.68**
2	DS@GT	gpt-4o	0.64	0.59	0.88	0.71	0.17	0.63	0.27	0.82	0.99	0.89	0.52	0.97	0.67
3	DS@GT	gpt-4.1	0.62	0.58	0.75	0.65	0.15	0.52	0.24	0.82	0.98	**0.90**	0.52	0.99	**0.68**
4	DS@GT	gemini-2.5-pro	0.62	0.59	0.67	0.63	0.17	0.52	0.25	0.84	0.97	**0.90**	0.52	0.98	**0.68**
5	SINAI	gritty-stock	0.56	0.60	0.60	0.60	0.19	0.40	0.25	0.84	0.86	0.85	0.50	0.57	0.53
6	DS@GT	claude-sonnet-4	0.56	0.56	0.43	0.49	0.15	0.36	0.21	0.83	0.92	0.88	0.51	0.93	0.66
7	SINAI	staff-frame	0.55	0.59	0.64	0.61	0.16	0.32	0.21	0.84	0.80	0.82	0.52	0.64	0.57
8	SINAI	radiant-tread	0.54	0.58	0.53	0.55	0.20	0.35	0.25	0.87	0.75	0.81	0.53	0.56	0.54
9	SINAI	iron-rhythm	0.52	0.57	0.46	0.51	0.15	0.37	0.21	0.84	0.79	0.81	0.50	0.63	0.56
10	DS@GT	claude-opus-4	0.51	0.49	0.21	0.29	0.16	0.31	0.21	0.85	0.90	0.88	0.51	0.92	0.66
11	SINAI	grating-dragster	0.49	0.59	0.63	0.61	0.20	0.58	0.30	0.84	0.39	0.53	0.50	0.54	0.52
12	SINAI	coped-message	0.39	0.57	0.32	0.41	0.17	0.21	0.19	0.84	0.67	0.74	0.45	0.16	0.24
13	SINAI	sizzling-coulomb	0.35	0.63	0.40	0.49	0.16	0.17	0.16	0.84	0.44	0.58	0.41	0.10	0.16

Retrieval-Augmented Debating (RAD)

Observations

- ❑ Some claims too hard to argue (e.g., the earth is flat).
  - Participant systems admitted defeat (*“you are right”*).
- ❑ Grounding responses in retrieved argument is hard.
  - Low quality score for most systems.
- ❑ LLMs do not recognize stance switches.
  - Systems pretended to disagree but argued for user stance.
- ❑ Common problem: wordiness.
  - Complex vocabulary, unclear argument, repetition.

→ Building a persuasive debate system is a hard task.