

Interactive Knowledge Assistance Track (iKAT) 2025 Overview

Organizers

Mohammad Aliannejadi

Simon Lupart

Marcel Gohsen

Zahra Abbasiantaeb

Nailia Mirzakhmedova

Johannes Kiesel

Jeff Dalton

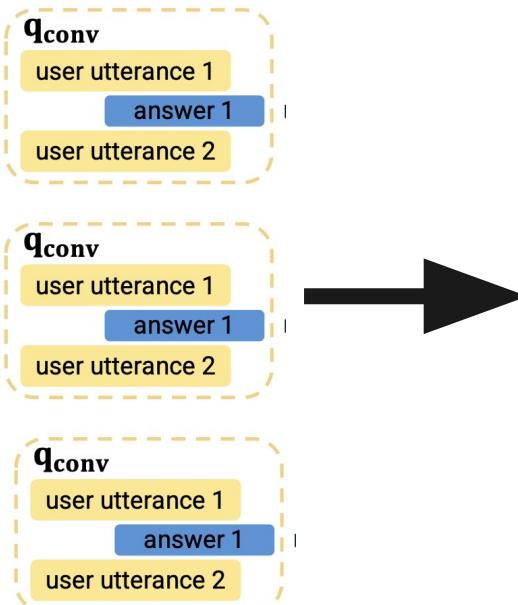


ChatGPT

Based on what you know about me, draw a picture of what you think my current life looks like



ChatGPT Memory



Memory

Simon has been added as an author to a new version of a paper on arXiv.



Simon is going to Washington DC in a week and is from Europe.



Simon fait de l'escalade.



Simon uses the LLMeval class with the SOLAR-107B model.



Simon is interested in scenarios for adversarial attacks where adding a page on the web could manipulate a system's rankings or retrieval results.



Simon is interested in the ClueWeb09 Part B dataset and has access to the IRLab server, where he is working with ClueWeb09 Part B data files stored on it.



L'utilisateur s'appelle Simon Lupart et est doctorant de 2ème année au IRLab.



L'utilisateur cuisine des pancakes et souhaite utiliser de la farine complète.



Effacer la mémoire

PTKB - Personal Text Knowledge Base



ChatGPT

iKAT Personal Textual Knowledge Base (PTKB): finding a university



I'm from the Netherlands.

I speak English fluently.

I don't have a driver's license.

I graduated from UvA.

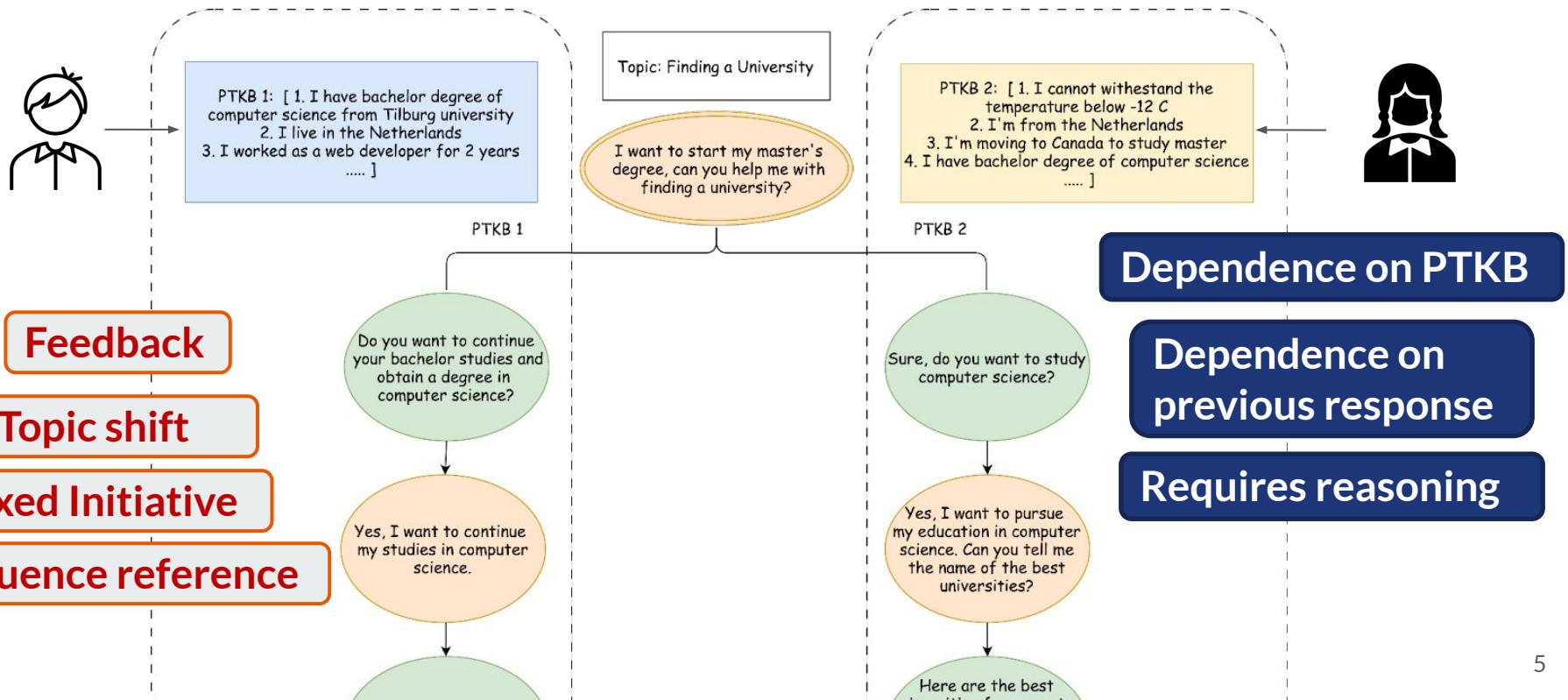
I'm used to heavy rains in the Netherlands.

I don't like crazy cold weather.

I plan to move to Canada.

I have bachelor's degree in computer science.

Using the Personal Textual Knowledge Base (PTKB)



iKAT Tasks Y3 – What's New

Three tasks for each turn (for each topic and each PTKB)

- **PTKB Binary Classification:** is this PTKB statement useful given the turn. **Evolving sets of PTKB statements across multiple dialogues.**
- **Passage Ranking:** provide a ranking of passages given the turn et context and PTKB.
 - a. **Dynamic pooling:** Going deeper in the pool with the help of LLMs
- **Response Generation:** provide a grounded and natural response given the turn et al.
 - a. **NoAnswer Turns:** Turns that do not have a correct answer – test LLMs generation.
 - b. **Evaluation:** Human-written responses.
 - c. **Nuggets-based Evaluation:**
 - Extracting Information Pieces from Passages (with Human).
 - Does the generated response covers every aspect (Prompting LLM).

iKAT Tasks Y3 – What's New

Interactive Response Generation Task

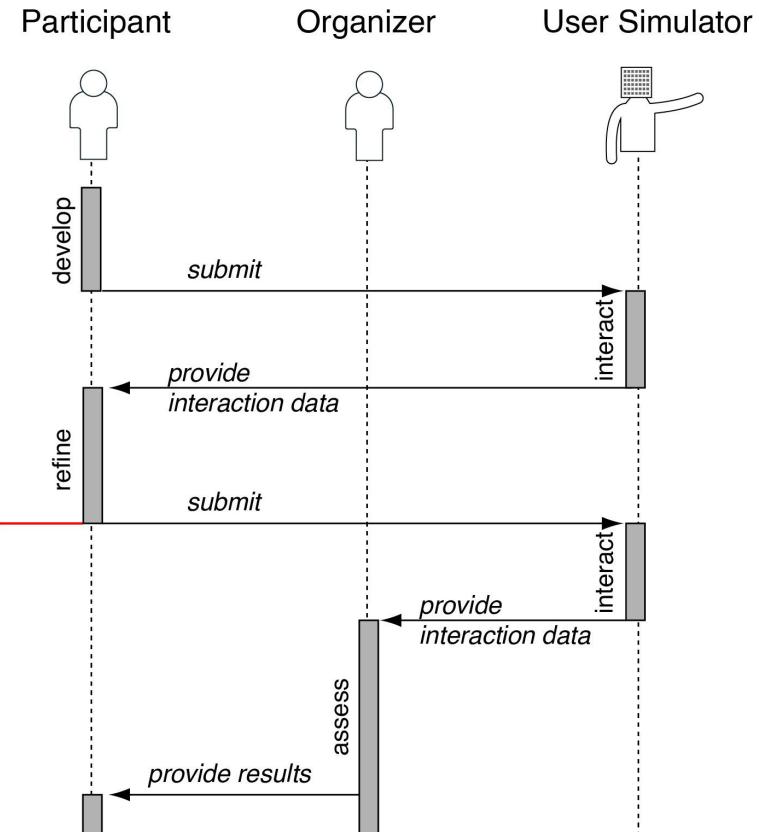
- Real-time conversations between simulated users and systems

Passage Ranking

- Providing rankings of relevant citations for generated responses

PTKB Extraction

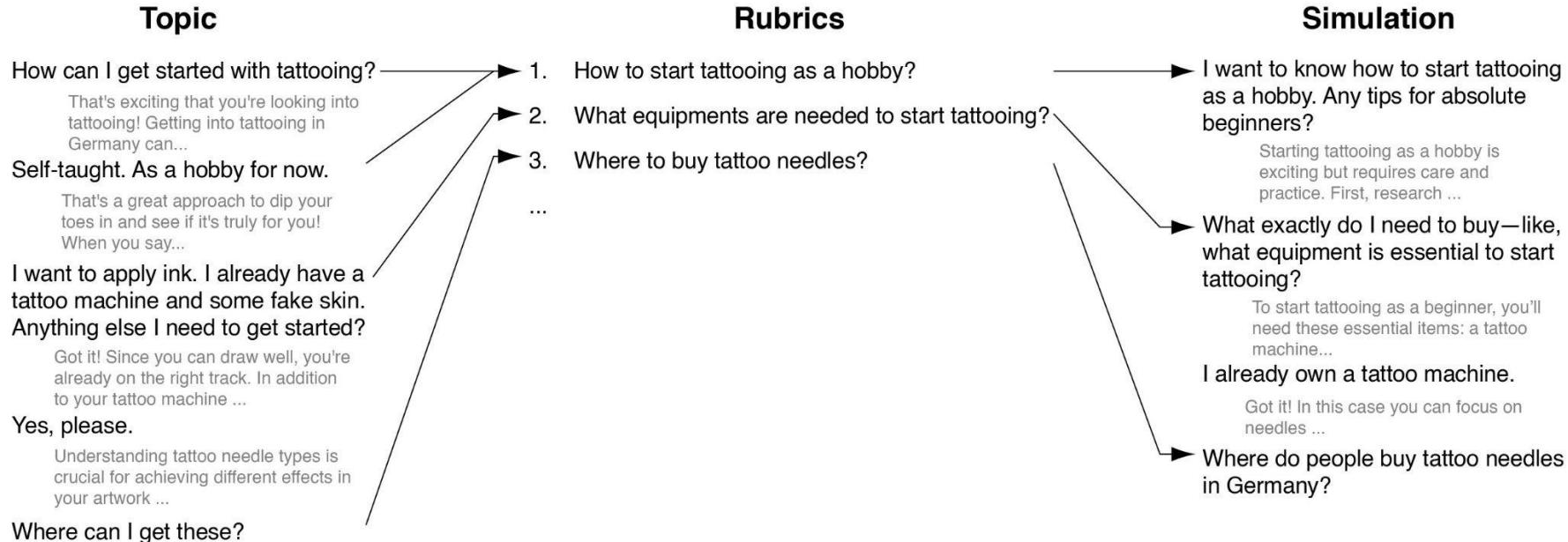
- Extracting revealed persona properties from conversations



iKAT Y3 Topics

- 17 conversational topics (12 assessed) – 188 turns (45 assessed).
- **Multifaceted Information Needs:** topics focused on exploring facets of broad topics.
 - e.g. “Tattooing 101”
 - e.g. “How to make good coffee”
- **Varied discourse:**
 - Revealment, feedback, clarifying questions, elicitation, etc..
- **Multiple personas** – personal textual knowledge base (**PTKB**).
 - 9 unique user personas (~20 statements each)
- **Multiple dialogs per persona:**
 - Reference prior conversations
 - Learn new relevant PTKBs of persona

User Simulation



iKAT Y3 Participation

- 7 teams
- 45 submitted runs + 14 baselines
- General trend
 - Breaking down the task for LLMs
- Three categories
 - automatic
 - interactive
 - generation-only (ranking provided)

Run ID	Type	Run ID	Type
<i>organizers</i>		<i>genaius</i>	
orga-bm25-human	auto	genaius-genonly-summary-gpt4o	gen_only
orga-splade-norerank	auto	genaius-genonly-full-gpt4o	gen_only
orga-splade-llama70b	auto	genaius-full-rewrite	interactive
orga-bm25-personal	auto	genaius-summary-rewrite	interactive
orga-ance-llama70b	auto		
orga-ance-norerank	auto		
orga-bm25-nopersonal	auto		
orga-gpt41mini-bm25-minilm-llama70b	interactive		
orga-gpt41mini-bm25-minilm-llama70b-nopersonal	interactive		
orga-llama8b-bm25-minilm-llama8b-v2	interactive		
orga-llama70b-bm25-minilm-llama70b	interactive		
orga-no-no-no-gpt41mini	interactive		
orga-no-no-no-llama8b-v2	interactive		
orga-no-no-no-llama70b	interactive		
<i>cfd</i>			
auto_npr1_npsg20_thru03_d3c5	auto		
auto_npr_npsg20_thru03_d3c5	auto		
auto_ori_npsg20_thru03_d3c5	auto		
auto_REnpr_npsg20_thru03_d3c5	auto		
gen-only_npsg13_thru0_d4c5	gen_only		
gen-only_npsg20_thru03_d3c5	gen_only		
cfda-adarewriter-chiq-llm4cs-splade	interactive		
cfda-chiq-llm4cs-splade-rff	interactive		
<i>grillab</i>			
grilllab-larf-finetuned-10-rounds-10-rounds	auto		
grilllab-larf-finetuned-10-rounds	auto		
grilllab-larf-finetuned	auto		
grilllab-larf-finetuned-rankilm	auto		
grilllab-larf-finetuned-22-rounds	auto		
cossine-orconvqa	auto		
agg_true-qrec-mse	auto		
agg_false-qrec-mse	auto		
grilllab-agentic-gpt4.1	interactive		
grilllab-agentic-gpt4.1-larf	interactive		
grilllab-agentic-gpt4.1-larf-v2	interactive		
grilllab-larf-fine-tuned-judge	interactive		

Evaluation

The evaluation metrics for each task:

- **PTKB Statement Binary Classification:**
 - Precision, Recall, F1 (using Organizer and NIST Assessor Judgements)
- **Passage Ranking:** Turn-level effectiveness
 - Primary Measure: nDCG@5
 - Other measures: MMR, MAP, etc.
 - Binary Relevance (threshold ≥ 2)
- **Response Generation:**
 - Human gold response and nugget set.
 - **Groundedness:** Is the response grounded by the passage ranking results (or nominated provenance results)?
 - **Nugget Recall:** Is the response, clear, concise and readable?
 - **RAG-based metrics:** LLMEval, BEM, etc.

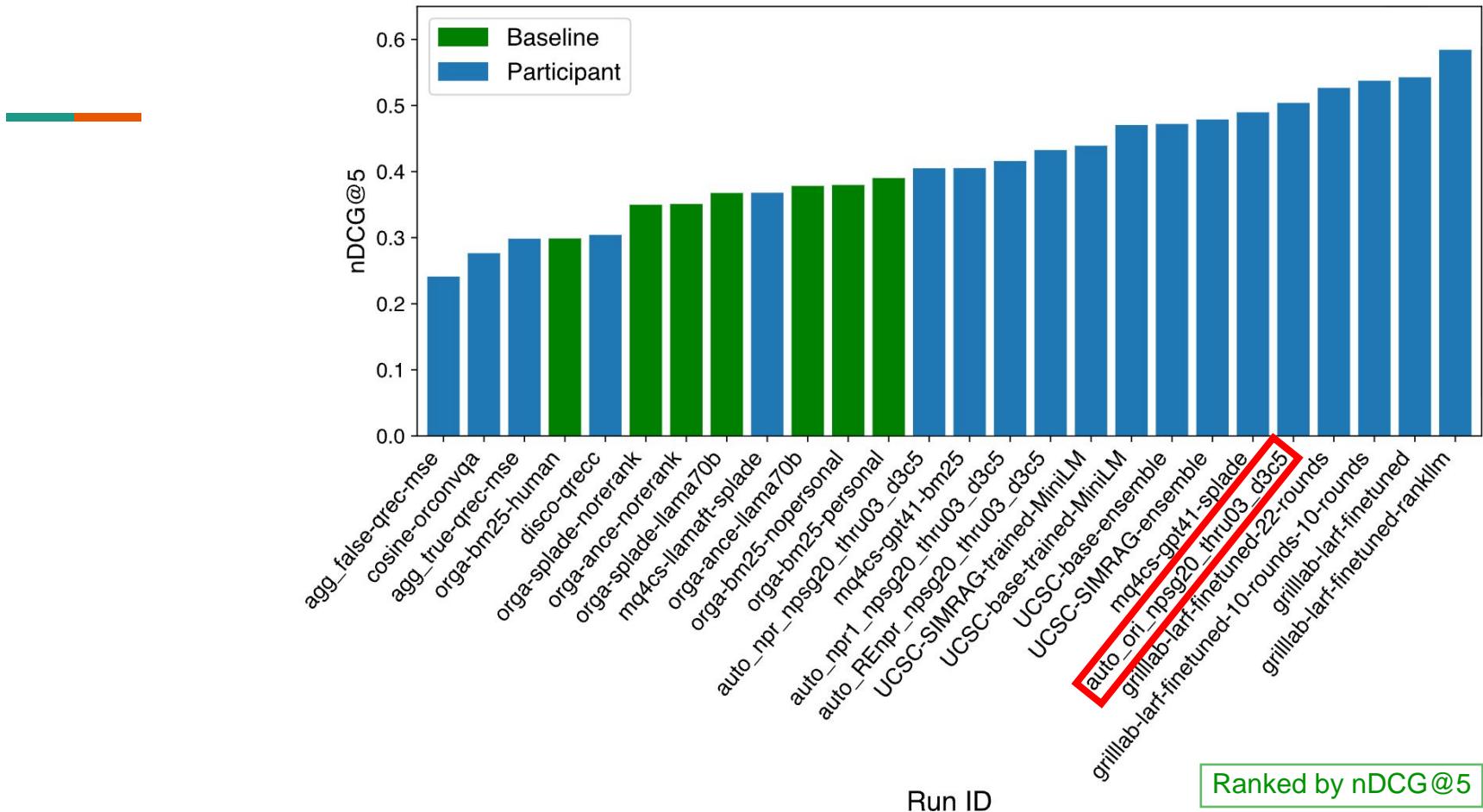
Evaluation (human assessment)

The evaluation metrics for interactive response generation:

- **Rubric level:**
 - **Engagement:** System encourages user to engage with it?
 - **Relevance:** Response are relevant to user's request?
 - **Overall Quality:** All factors considered, how good are the responses?
- **Dialogue level:**
 - **Mixed-initiative:** System uses proactive methods?
 - **Personalization:** System personalize its responses to the user?
 - **Information-flow:** System uses information scaffolding?
 - **Trustworthiness:** System appears to be trustworthy?
 - **User Satisfaction:** System appears to satisfy the user?

Results

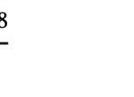
Automatic Passage Ranking Results



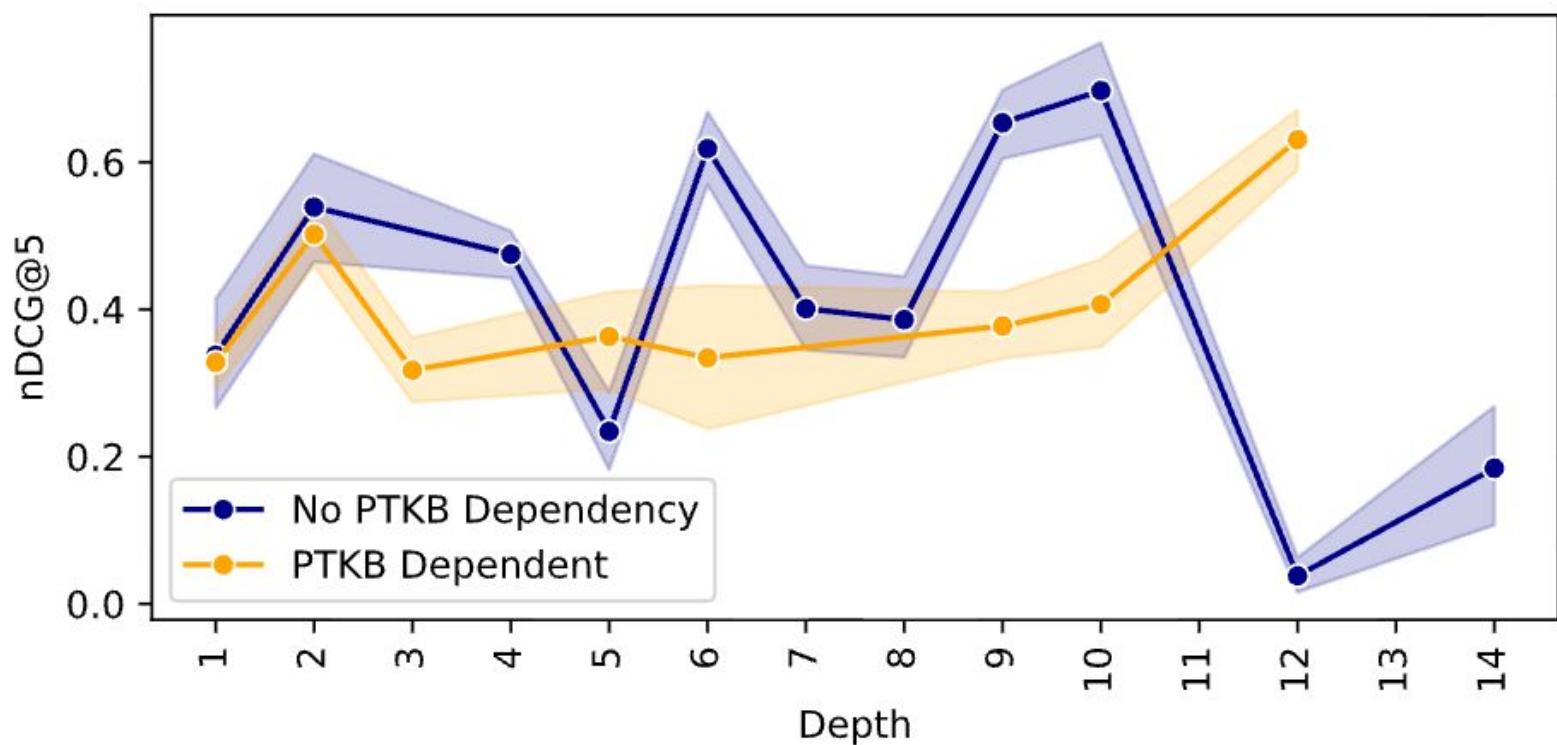
Automatic Passage Ranking Results

Group	Run ID	nDCG@3	nDCG@5	nDCG	P@20	Recall@20	Recall	mAP
grilllab	grilllab-larf-finetuned-rankllm	0.5941	0.5843	0.5070	0.5867	0.2173	0.5273	0.3041
	grilllab-larf-finetuned	0.5489	0.5427	0.4991	0.5711	0.2137	0.5245	0.2943
	grilllab-larf-finetuned-10-rounds-10-rounds	0.5394	0.5375	0.5025	0.5711	0.2130	0.5334	0.2953
	grilllab-larf-finetuned-22-rounds	0.5346	0.5267	0.5048	0.5656	0.2134	0.5412	0.2955
cfda	auto ori npsg20 thru03 d3c5	0.5239	0.5040	0.4116	0.5522	0.2037	0.3721	0.2375
uva	mq4cs-gpt41-splade	0.4916	0.4897	0.5357	0.5789	0.2073	0.6101	0.3099
ucsc	UCSC-SIMRAG-ensemble	0.4787	0.4789	0.5462	0.5656	0.2033	0.6648	0.3232
ucsc	UCSC-base-ensemble	0.4816	0.4721	0.5624	0.5622	0.2019	0.7016	0.3252
ucsc	UCSC-base-trained-MiniLM	0.4562	0.4705	0.5264	0.4544	0.1677	0.7016	0.2551
ucsc	UCSC-SIMRAG-trained-MiniLM	0.4502	0.4392	0.5116	0.4800	0.1758	0.6648	0.2592
cfda	auto_REnpr_npsg20_thru03_d3c5	0.4458	0.4326	0.3731	0.5056	0.1778	0.3619	0.2133
cfda	auto_npr1_npsg20_thru03_d3c5	0.4153	0.4160	0.3638	0.4900	0.1786	0.3588	0.2046
uva	mq4cs-gpt41-bm25	0.4180	0.4053	0.4126	0.4756	0.1675	0.4481	0.2139
cfda	auto_npr_npsg20_thru03_d3c5	0.4135	0.4050	0.3593	0.4900	0.1775	0.3546	0.2091
organizers	orga-bm25-personal	0.4074	0.3902	0.3779	0.4178	0.1498	0.4308	0.1871
organizers	orga-bm25-nopersonal	0.3739	0.3797	0.3779	0.4378	0.1572	0.4307	0.1950
organizers	orga-ance-llama70b	0.3937	0.3783	0.4245	0.4044	0.1427	0.5267	0.2187
uva	mq4cs-llamaft-splade	0.3612	0.3680	0.3911	0.4756	0.1701	0.4413	0.2045
organizers	orga-splade-llama70b	0.3753	0.3676	0.4816	0.4122	0.1470	0.6721	0.2320
organizers	orga-ance-norerank	0.3692	0.3510	0.3872	0.3989	0.1414	0.5258	0.1714
organizers	orga-splade-norerank	0.3555	0.3498	0.4719	0.4000	0.1545	0.6760	0.2150
uva	disco-qrec	0.3087	0.3042	0.3926	0.3544	0.1347	0.5524	0.1683
organizers	orga-bm25-human	0.3021	0.2988	0.2955	0.3411	0.1336	0.3514	0.1492
guidance	agg_true-qrec-mse	0.3161	0.2985	0.3692	0.3167	0.1233	0.5185	0.1690
guidance	cosine-orconvqa	0.2715	0.2765	0.3609	0.3233	0.1225	0.5205	0.1549
guidance	agg_false-qrec-mse	0.2483	0.2411	0.3409	0.2922	0.0943	0.5140	0.1378

Ranked by nDCG@3



Passage Ranking Results by Conversation Depth



Generated Response Evaluation

Group	Run ID	LLM eval		Nugget recall	BEM	F1	ROUGE-1
		SOLAR	GPT-4.1				
NIST	gold-human-response	-	-	0.2509	-	-	-
grilllab	grilllab-larf-finetuned-22-rounds	0.9111	0.7778	0.2507	0.1759	0.2858	0.2103
grilllab	grilllab-larf-finetuned	0.9778	0.8000	0.2418	0.1715	0.2887	0.2138
grilllab	grilllab-larf-finetuned-10-rounds-10-rounds	0.9556	0.8000	0.2203	0.1887	0.2830	0.2079
ucsc	UCSC-base-ensemble	0.9091	0.8222	0.1902	0.1682	0.3174	0.2528
ucsc	UCSC-SIMRAG-trained-MiniLM	0.8889	0.7556	0.1870	0.1717	0.3136	0.2555
ucsc	UCSC-base-trained-MiniLM	0.8889	0.8222	0.1867	0.1990	0.3126	0.2530
ucsc	UCSC-SIMRAG-ensemble	0.9333	0.8444	0.1787	0.1836	0.3168	0.2538
grilllab	grilllab-larf-finetuned-rankllm	0.9556	0.8222	0.1768	0.1649	0.2871	0.2120
uva	mq4cs-llamaft-splade	0.9111	0.7778	0.1510	0.1513	0.3038	0.2561
uva	mq4cs-gpt41-splade	0.9111	0.8444	0.1490	0.1822	0.3134	0.2585
uva	mq4cs-gpt41-bm25	0.8444	0.8222	0.1423	0.1616	0.2981	0.2501
uva	disco-qrec	0.8667	0.8222	0.1308	0.1468	0.3068	0.2595
cfda	auto_ori_npsg20_thru03_d3c5	0.9111	0.7333	0.1145	0.1444	0.2990	0.2541
guidance	cosine-orconvqa	0.7778	0.6667	0.1111	0.1552	0.2707	0.2291
organizers	orga-bm25-nopersonal	0.8000	0.6889	0.1099	0.1507	0.2937	0.2500
guidance	agg_true-qrec-mse	0.7556	0.7111	0.1073	0.1426	0.2804	0.2324
guidance	agg_false-qrec-mse	0.8000	0.6444	0.1041	0.1612	0.2713	0.2299
cfda	auto_REnpr_npsg20_thru03_d3c5	0.9556	0.8000	0.1018	0.1626	0.2929	0.2481
cfda	auto_npr1_npsg20_thru03_d3c5	0.9556	0.7778	0.0991	0.1516	0.2869	0.2467
cfda	auto_npr_npsg20_thru03_d3c5	0.9111	0.8222	0.0972	0.1479	0.2907	0.2471
organizers	orga-bm25-personal	0.7727	0.6222	0.0924	0.1849	0.3110	0.2676
organizers	orga-ance-norerank	0.8222	0.6222	0.0863	0.1702	0.2969	0.2580
organizers	orga-splade-llama70b	0.7333	0.6667	0.0830	0.1497	0.3010	0.2651
organizers	orga-bm25-human	0.8222	0.6222	0.0780	0.1813	0.2966	0.2645
organizers	orga-splade-norerank	0.7556	0.5556	0.0614	0.1565	0.2838	0.2425
organizers	orga-ance-llama70b	0.7778	0.6889	0.0576	0.1523	0.2995	0.2625
<i>Generation only</i>							
cfda	gen-only_npsg13_thru0_d4c5	0.9333	0.8000	0.1195	0.1641	0.3136	0.2689
uva	nuggets-noptkb	0.9111	0.8222	0.1070	0.1721	0.3026	0.2537
guidance	genonly_clariftop10	0.7778	0.6889	0.1041	0.1650	0.2799	0.2306
uva	nuggets-ptkb	0.8667	0.7778	0.1030	0.1923	0.3052	0.2524
genaius	genaius-genonly-full-gpt4o	0.8889	0.7556	0.0999	0.1672	0.2827	0.2485
cfda	gen-only_npsg20_thru03_d3c5	0.9111	0.7778	0.0978	0.1524	0.3065	0.2552
genaius	genaius-genonly-summary-gpt4o	0.8667	0.7556	0.0811	0.1407	0.2750	0.2500
usiir	usiir_run2	0.4000	0.2222	0.0510	0.1267	0.1877	0.1708
usiir	usiir_run1	0.6000	0.3111	0.0508	0.1186	0.1916	0.1740

Retrieve-then-respond
(ranked by nugget recall)

Generation only
(ranked by nugget recall)

Interactive Response Generation Evaluation

Table 7: Evaluation results of runs in the interactive task based on human assessments. On the rubric level, assessor evaluated engagement (Eng), relevance (Rel), quality (Qual), and their confidence in the ratings (Conf). On the dialogue level, assessors rated mixed-initiative strategies (Mix), personalization (Pers), information flow (Flow), trustworthiness (Trust), user satisfaction (Sat), and their confidence in these ratings (Conf).

Run ID	Ranked by Avg. Score	Rubric Level					Dialogue Level						Score	
		Eng	Rel	Qual	Conf	Score	Mix	Pers	Flow	Trust	Sat	Conf		
orga-no-no-no-gpt41mini		0.81	0.85	0.85	0.96	0.84	0.45	0.92	0.97	0.47	0.82	0.94	0.78	0.79
genaius-full-rewrite		0.71	0.80	0.82	0.94	0.78	0.46	0.59	0.93	0.72	0.82	0.94	0.78	0.77
genaius-summary-rewrite		0.72	0.80	0.85	0.95	0.81	0.48	0.59	0.91	0.76	0.78	0.90	0.72	0.75
cfda-adarewriter-chiq-llm4cs-splade		0.64	0.81	0.84	0.94	0.78	0.33	0.46	0.91	0.72	0.81	0.92	0.75	0.74
cfda-chiq-llm4cs-splade-rff		0.63	0.79	0.81	0.95	0.77	0.31	0.41	0.94	0.71	0.76	0.94	0.72	0.74
grilllab-agentic-gpt4.1-larf		0.67	0.73	0.74	0.89	0.68	0.64	0.55	0.78	0.82	0.76	0.92	0.72	0.69
uva-gpt5mini-bm25-debertav3-gpt5mini		0.65	0.73	0.73	0.95	0.71	0.42	0.57	0.74	0.65	0.76	0.86	0.69	0.68
grilllab-agentic-gpt4.1-larf-v2		0.71	0.75	0.79	0.90	0.73	0.81	0.62	0.82	0.76	0.72	0.84	0.64	0.67
uva-gpt5mini-no-no-gpt5mini		0.62	0.72	0.72	0.92	0.68	0.46	0.57	0.74	0.49	0.72	0.88	0.66	0.66
orga-llama8b-bm25-minilm-llama8b-v2		0.60	0.68	0.70	0.94	0.67	0.25	0.37	0.66	0.63	0.72	0.94	0.68	0.66
uva-gpt5-bm25-debertav3-gpt5mini-nopersonal		0.51	0.66	0.67	0.92	0.63	0.33	0.24	0.68	0.66	0.71	0.92	0.67	0.64
grilllab-agentic-gpt4.1		0.68	0.69	0.74	0.92	0.70	0.65	0.54	0.84	0.81	0.69	0.88	0.63	0.63
ucsc-SIMRAG-keywordQuery-dynamicPTKB-ensembleReranker		0.58	0.65	0.67	0.97	0.67	0.24	0.38	0.79	0.66	0.63	0.96	0.61	0.62
ucsc-SIMRAG-guidelineQuery-dynamicPTKB-trainedReranker		0.54	0.60	0.63	0.93	0.61	0.25	0.40	0.74	0.69	0.68	0.94	0.64	0.62
ucsc-SIMRAG-keywordQuery-dynamicPTKB-trainedReranker		0.58	0.64	0.66	0.93	0.64	0.33	0.37	0.88	0.69	0.66	0.94	0.63	0.61
orga-gpt41mini-bm25-minilm-llama70b		0.58	0.65	0.65	0.92	0.62	0.42	0.50	0.74	0.63	0.66	0.94	0.63	0.61
ucsc-base-dynamicPTKB-trainedReranker		0.55	0.60	0.62	0.95	0.59	0.22	0.42	0.74	0.63	0.62	0.92	0.58	0.57
orga-no-no-no-llama70b		0.63	0.63	0.63	0.90	0.59	0.25	0.56	0.68	0.44	0.57	0.86	0.52	0.53
uva-gpt5-bm25-debertav3-gpt5		0.44	0.58	0.58	0.94	0.55	0.19	0.43	0.44	0.60	0.57	0.90	0.52	0.52
orga-llama70b-bm25-minilm-llama70b		0.54	0.59	0.59	0.92	0.56	0.37	0.54	0.60	0.66	0.60	0.82	0.52	0.51
orga-no-no-no-llama8b-v2		0.56	0.61	0.61	0.89	0.57	0.36	0.54	0.63	0.47	0.59	0.80	0.49	0.50
orga-gpt41mini-bm25-minilm-llama70b-nopersonal		0.52	0.55	0.58	0.90	0.53	0.27	0.23	0.62	0.50	0.47	0.84	0.39	0.42
grilllab-larf-fine-tuned-judge		0.49	0.49	0.54	0.74	0.44	0.33	0.59	0.62	0.74	0.60	0.59	0.36	0.38

(Planning) iKAT Y4 2026

- Continue beyond TREC :-)
- Simulation of diverse search behavior
 - Testing systems under various simulation scenarios
- Focus on generation for complex tasks (including rationales)
- Multiple conversations per persona

Thank you

