CFDA & CLIP at TREC iKAT 2025: Enhancing Personalized Conversational Search via Query Reformulation and Rank Fusion

1 OVERVIEW AT A GLANCE

TREC iKAT (Interactive Knowledge Assistance Track) evaluates personalized conversational information retrieval systems. We submitted systems to both TREC iKAT 2025 evaluation tracks:

- **Online Track:** 2 systems evaluated on human assessment (13 metrics) and retrieval (6 metrics)
- **Offline Track:** 4 systems evaluated on document retrieval and PTKB[1] statement ranking

Table 1: Overall Performance Summary

Track	Task / Metric	Best	Med	Max	vs Med
Online (new)	Human: Total Score	74.0	63.0	79.0	+11.0
	Retrieval: nDCG@3	17.1	11.3	23.5	+5.8
	Retrieval: nDCG@5	17.0	10.4	22.9	+6.6
Offline	Doc: nDCG@5	50.7	40.7	74.0	+10.0
	PTKB: nDCG@5	38.8	42.6	75.0	-3.8

Online Systems: On#1 (AdaRewriter) uses adaptive query reformulation with SPLADE; **On#2 (RRF)** applies Reciprocal Rank Fusion. **Offline Systems: Off#1** (adaptive rewriting), **Off#2** (expanded reformulation), **Off#3** (passage scoring), **Off#4** (hybrid).

2 ONLINE SUBMISSION RESULTS

2.1 Human Evaluation

Human assessors evaluate response quality across dimensions:

- **Coherence/Relevance:** Does the system maintain logical flow and provide useful answers?
- **Personalization:** Does the system leverage user preferences from PTKB?
- **Mixed-Initiative:** Does the system proactively ask clarifying questions?

Table 2: Online: Human Evaluation Results

	Ours		Competition		
Metric	**On#1**	**On#2**	**Med**	**Max**	**vs Med**
Coherence	0.91	**0.94**	0.74	0.97	+0.20
Relevance	**0.81**	0.79	0.66	0.85	+0.15
Satisfaction	**0.81**	0.76	0.69	0.82	+0.12
Personalization	**0.46**	0.41	0.54	0.92	-0.08
Mixed-Initiative	**0.33**	0.31	0.33	0.81	0
Total Score	**0.74**	**0.74**	0.63	0.79	+0.11

2.2 Retrieval Evaluation

- **nDCG@k:** Ranking quality at top-k positions
- **P@k/Recall@k:** Precision and coverage at cutoff k

[1]Personal Text Knowledge Base (PTKB): a set of natural language statements describing user characteristics, preferences, and constraints (e.g., "I am lactose intolerant") that enables personalized responses.

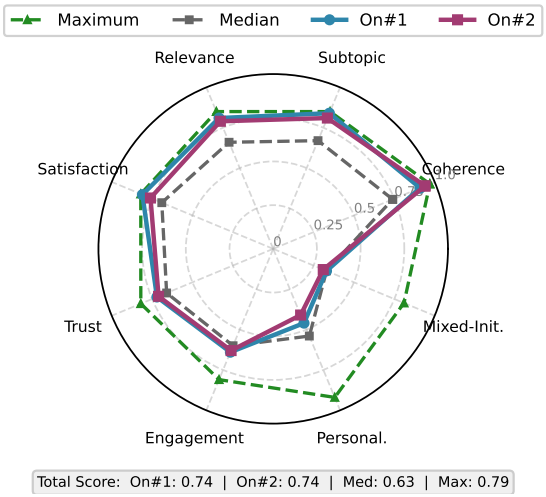

Figure 1: Human evaluation comparison.

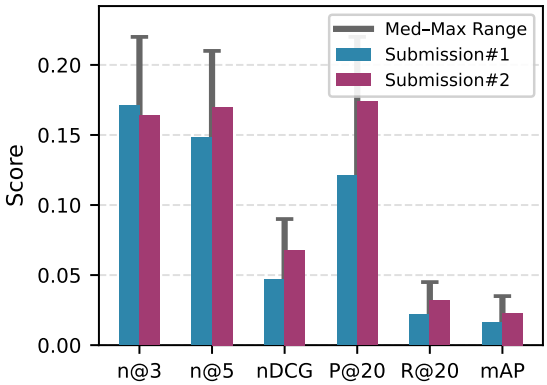

Figure 2: Retrieval metrics comparison. On#1 excels at early precision (nDCG@3), On#2 provides better depth coverage.

Table 3: Online: Retrieval Results

	Ours		Competition		
Metric	**On#1**	**On#2**	**Med**	**Max**	**vs Med**
nDCG@3	**17.1**	16.4	11.3	23.5	+5.8
nDCG@5	14.8	**17.0**	10.4	22.9	+6.6
nDCG	4.7	**6.8**	2.9	29.1	+3.9
P@20	12.1	**17.4**	4.8	26.7	+12.6
Recall@20	2.2	**3.2**	1.3	6.8	+1.9
mAP	1.6	**2.3**	1.1	9.6	+1.2

3 OFFLINE SUBMISSION RESULTS

Offline evaluation uses automatic metrics on held-out test data:

- **Document Retrieval:** Measures ability to find relevant passages from the corpus
- **PTKB Retrieval:** Measures ability to identify relevant user preference statements

3.1 Document Retrieval

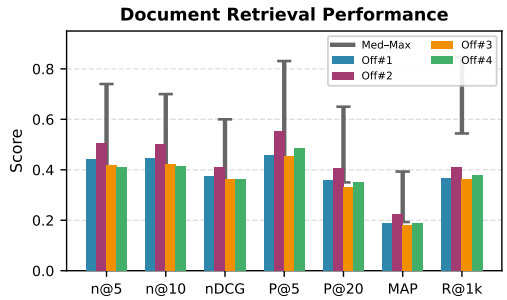

Figure 3: Document retrieval comparison. Gray bars show Median-to-Maximum range.

Table 4: Offline: Document Retrieval Results

	Ours				Competition		
Metric	**Off#1**	**Off#2**	**Off#3**	**Off#4**	**Med**	**Max**	**vs Med**
nDCG@5	44.3	**50.7**	41.9	41.1	40.7	74.0	+10.0
nDCG@10	44.7	**50.0**	42.3	41.5	41.3	71.2	+8.7
P@5	45.8	**55.1**	45.3	48.4	44.4	83.1	+10.7
MAP	19.0	**22.4**	18.1	18.8	19.3	39.3	+3.1
Recall@1000	36.8	**41.1**	36.4	38.0	54.4	84.7	-13.3

3.2 PTKB Retrieval

Table 5: Offline: PTKB Retrieval Results

	Ours				Competition		
Metric	**Off#1**	**Off#2**	**Off#3**	**Off#4**	**Med**	**Max**	**vs Med**
nDCG@5	**38.8**	33.9	35.7	35.7	42.6	75.0	-3.8
nDCG@10	**34.8**	30.1	31.7	32.2	39.2	64.8	-4.4
P@5	**28.0**	24.4	26.2	25.3	35.1	62.2	-7.1
MAP	**22.2**	18.3	20.0	20.4	26.2	56.2	-4.0
Recall@1000	**22.4**	19.3	20.2	20.6	27.3	63.0	-4.9

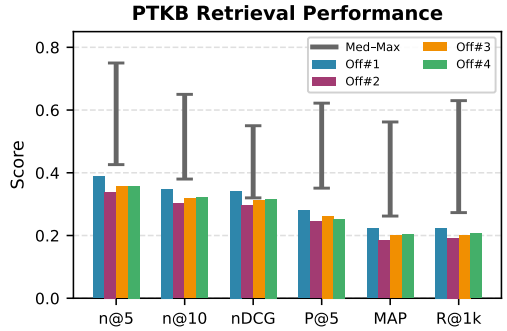

Figure 4: PTKB retrieval comparison. All metrics below competition median.

4 CONCLUSION AND FUTURE WORK

Our systems achieved strong performance in coherence (+0.20) and retrieval (+0.058 nDCG@3), but underperformed in personalization and PTKB retrieval. Future directions:
Online: Enhance mixed-initiative dialogue and personalization by better integrating PTKB statements into response generation.
Offline: Improve PTKB retrieval through user preference modeling and explore multi-turn context for document ranking.