

SemEval-2026 Task 8: MTRAGEval: Evaluating Multi-Turn RAG Conversations

Sara Rosenthal, Vraj Shah, Yannis Katsis, Marina Danilevsky

IBM Research, USA
sjrosenthal@us.ibm.com

Abstract

We present the results and findings from SemEval Task 8: MTRAGEval. MTRAGEval measures three Retrieval Augmented Generation (RAG) subtasks: A. Retrieval, B. Generate, and C. Retrieve+Generate (full RAG) on multi-turn conversations. The task is evaluated using MTRAG-UN, a new benchmark for Multi-Turn RAG focusing on Unanswerable, Underspecified, Non-Standalone, and Unclear Questions. The MTRAGEval task attracted strong participation with 107 registered teams and 92 submissions across all subtasks, and yielded several interesting findings on effective retrieval and query rewriting techniques, the use of ensemble models, and the compounding costs of retrieval errors on downstream generation quality.

1 Introduction

Real-world conversational AI frequently depends on Retrieval-Augmented Generation (RAG) techniques to deliver reliable, grounded responses. Question answering research has sought to address these challenges, as seen for instance in the recent TREC RAG eval (Pradeep et al., 2024) shared task on single questions. Work has also moved beyond single question answering to multi-turn (MT) conversations (Aliannejadi et al., 2024; Dziri et al., 2022; Feng et al., 2021; Kuo et al., 2025; Katsis et al., 2025). This includes the TREC iKAT (Aliannejadi et al., 2024) task, which focused on personalized MTRAG. For further related work on MTRAG benchmarks refer to our benchmark papers (Katsis et al., 2025; Rosenthal et al., 2026).

In real conversations, questions can reference prior context, be unanswerable or change topics — challenges that make multi-turn RAG substantially harder to evaluate than single-turn settings. In particular, the MTRAG benchmark (Katsis et al., 2025) has shown that there is still plenty of room for improvement in RAG-based chat. It is the first benchmark that uses active (i.e., real-time) retrieval,

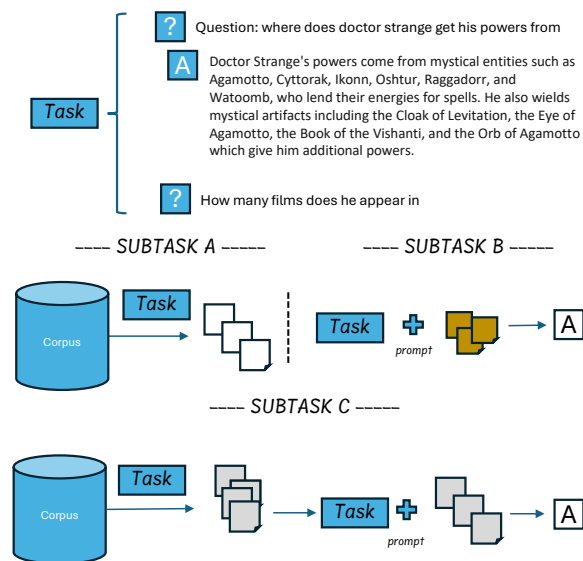


Figure 1: A description of the input and task for the three proposed subtasks. A Conversation task with two turns is shown. Subtask A) Given a Conversation task, Retrieve the relevant passages for the last turn in the task. Subtask B) Given a Conversational task and the reference passages (in gold), generate an answer to the last turn in the task. Subtask C) Given a Conversational task, retrieve passages and then generate an answer to the last turn in the task. Input provided to participants is shown in blue (and gold), model output that will be evaluated is in white, and intermediate input is in grey.

includes unanswerable queries, and has multiple domains. These properties make it a good benchmark for evaluating all aspects of the RAG pipeline: A) Retrieval only, B) Generation only, and C) end-to-end RAG. This SemEval task expands on the MTRAG work by presenting three subtasks mirroring these aspects (see Figure 1):

- **Subtask A - Retrieval**
- **Subtask B - Generate on Reference Passages**
- **Subtask C - Retrieve and Generate (RAG)**

We provide the MTRAG dataset as training/validation and release the new MTRAG-UN

benchmark (Rosenthal et al., 2026) as a hidden evaluation set. MTRAG-UN contains 666 conversational tasks, including Unanswerable, Underspecified (surprise type), Non-Standalone, and Unclear questions. 507 of those conversational tasks were used in MTRAGEval.

The task yielded interesting findings across all three subtasks. In Subtask A, top teams used sparse or hybrid retrieval approaches, and consistently did query rewriting to resolve coreferences and produce self-contained queries prior to retrieval, reinforcing findings from prior work on multi-turn retrieval (Katsis et al., 2025). In Subtask B, more than half of participating teams exceeded the strongest single-LLM baseline, with many top teams employing ensembles of multiple model families for both generation and answer judgment. Notably, the vast majority of teams relied on zero-shot prompting, suggesting room for more sophisticated approaches. Comparing Subtask B and Subtask C reveals the compounding cost of retrieval errors on generation: the top H-Mean score dropped by approximately 20 points between the two subtasks, directly validating the importance of evaluating generation both with and without gold passages. Subtask C also saw a shift toward non-open-source models relative to Subtask B, although zero-shot prompting remained dominant across both subtasks.

MTRAGEval attracted strong participation, with 194 members on the mailing list and 107 registered teams. Across the three subtasks, a total of 48 teams submitted results — 38 (Subtask A), 26 (Subtask B), and 29 (Subtask C) — with some teams participating in multiple subtasks. The workshop received 31 paper submissions, reflecting broad interest in both system development and research questions surrounding multi-turn RAG evaluation.

2 Task Definition

Given a conversation, a task is a conversation turn containing all previous turns together with the last user question (e.g., the task created for turn k includes all user and agent questions/responses for the first $k - 1$ turns plus the user question for turn k). An example of a conversational task with two turns is shown in Figure 1. All of our subtasks are performed at the conversational task level. We have the following three subtasks (See Figure 1): A) Retrieval, B) Generation with reference passages, and C) full RAG - retrieval followed by generation. Participants were able to participate in one or multiple

subtasks. We encouraged creative submissions that use different resources such as new trained models, prompt engineering, query rewriting, and agentic RAG (repeated querying). We did not allow submissions using Mixtral 8x7B as this model was used as the generator during dataset creation.

2.1 Subtask A - Retrieval

Given a conversational task that is answerable and has a set of relevant passages, we evaluate how successful the retrieval system is at retrieving the relevant passages.

Input: A set of conversational tasks, where each task contains (a) a conversation of user question/agent response turns, ending with a user question and (b) the document corpus.

Output: For each conversational task, return an ordered list of 10 passages from the document corpus that are relevant to the last user question (with more relevant passages appearing earlier in the list). Note that the submission for these tasks were only evaluated on the subset of answerable and partial questions; however, to avoid information leak for the other subtasks, all conversational tasks were provided without info on which conversational tasks were answerable or not.

2.2 Subtask B - Generation with Reference Passages (Reference)

Given a conversational task with a set of relevant passages and a target answer, we evaluate the quality of the generated answer.

Input: A set of conversational tasks, where each conversational task contains (a) a conversation comprising of a set of user question/agent response turns ending with a user question and (b) a set of relevant passages for the last user question.

Output: For each conversational task, generate an agent response for the last user question (which should be faithful w.r.t. the relevant passages).

2.3 Subtask C - Generation with Retrieved Passages (RAG)

Given a conversational task with a target answer, the RAG system must retrieve passages and generate an answer. We evaluate the quality of the generated answer.

Input: A set of conversational tasks, where each task contains (a) a conversation of user question/agent response turns ending with a user question and (b) the document corpus.

	Recall		nDCG	
	@5	@10	@5	@10
BM25	0.20	0.27	0.18	0.21
BGE-base 1.5	0.30	0.38	0.27	0.30
Elser	0.49	0.58	0.45	0.49

Table 1: Retrieval Performance baselines of models on our benchmark metrics with the last turn question only. Retrieval performance is only computed on the 777 answerable and partial tasks of MTRAG

	RL _F		RB _{llm}		RB _{alg}	
	●	○	●	○	●	○
Reference	0.87	0.65	0.95	0.95	0.88	0.85
GPT-4o	0.76	<u>0.71</u>	0.76	0.70	<u>0.45</u>	0.40
Llama-3.1-405B	<u>0.75</u>	0.72	<u>0.74</u>	<u>0.68</u>	0.48	0.42
Llama-3.1-8B	0.55	0.56	0.59	0.59	0.37	0.35
Qwen2.5-7B	0.68	0.67	0.66	<u>0.68</u>	0.44	0.39

Table 2: Generation results on the 842 tasks: Reference (●) and RAG (○), w/ IDK conditioned metrics. Per column, the best result in **bold** and 2nd best is underlined.

Output: For each conversational task, first retrieve up to 10 passages from the documents corpus that are relevant to the user question and use them to generate a faithful agent response for the last user question.

3 Data

The MTRAGEval task uses the MTRAG (Katsis et al., 2025) and MTRAG-UN (Rosenthal et al., 2026) benchmarks for training/validation and testing. Examples showing multi-turn conversations are provided in each paper. The conversations are all manually created and reviewed by human annotators to ensure high quality using the RAGAPHENE tool (Fadnis et al., 2025). The annotators are highly skilled individuals whose sole job is to perform annotation tasks and they are paid well above minimum wage. Annotator information is anonymized to avoid PII. Further, the questions in the benchmark and test data are general and not specific to any individual. Any mention of information that looks personal is fictitious (e.g. How can I avoid bankruptcy?). All data is released under the Apache 2.0 license and is available at <https://github.com/IBM/mt-rag-benchmark>.

3.1 Training Data and Baselines

We released the MTRAG benchmark (Katsis et al., 2025) as trial/training data along with scripts for evaluation. The benchmark contains 110 conver-

sations which is 842 conversational tasks. We provided baselines in Table 1 and Table 2 respectively for retrieval and generation which showed there is room for improvement for all subtasks.

3.2 Test Data

The MTRAGEval test data is presented in the MTRAG-UN benchmark (Rosenthal et al., 2026). MTRAG-UN comprises human-generated multi-turn conversations with an average of 8 turns per conversation. The benchmark introduces five key conversational task types evaluated across multi-turn contexts: **Unanswerable Questions** are questions that cannot be answered from retrieved passages because no relevant passages exist. At least 2 unanswerable questions are included per conversation to ensure diverse coverage. **Underspecified Questions** are ambiguous, ill-formed, or lack sufficient information to determine a clear user intent. Rather than providing incorrect answers, systems should detect the ambiguity and either request clarification, present plausible interpretations, or list options based on available passages. These were created through a combination of human annotation and synthetic generation with query expansion methods. **Non-standalone Questions** appear in later conversation turns and implicitly reference information from earlier turns, requiring understanding of prior context and challenging retrieval systems to handle context-dependent queries. **Unclear Responses or Clarifications** are follow-up questions where users seek clarification because they did not understand or disagreed with the model’s previous response, and these are explicitly called out and evaluated.

From each conversation, a single turn was selected to create evaluation tasks, with preference given to challenging turns. The resulting conversational tasks show higher representation of unanswerable/partially answerable questions and include explicitly labeled clarification questions. The benchmark is also biased against selecting first turns, as these were found to be easier for language models. The benchmark includes 666 tasks across 6 domains. In MTRAGEval, we only keep the original MTRAG domains (CLAPNQ, FiQA, Govt, Cloud) consisting of 507 tasks: 285 answerable, 47 partially answerable, 97 unanswerable, and 78 underspecified. The answerable and partial tasks are used for the retrieval Subtask A and all conversational tasks are used for Subtask B and C, but the underspecified are presented as a surprise challenge

and not included in the leaderboard. For complete details on benchmark design and evaluation across all domains, refer to (Rosenthal et al., 2026).

4 Participating Teams

MTRAGEval had 48 teams participating in total. Out of them, 17 teams (35%) participated in all three subtasks, with the remaining teams submitting solutions to a subset. Among subtasks, Subtask A was the most popular with 38 teams participating (79%), with Subtask B and C having 26 (54%) and 29 (60%) teams, respectively. Table 8 in the Appendix shows detailed team information.

We shared a questionnaire with participants regarding participation and system descriptions as well as overall impression of the task which was completed by 45/48 teams. Participants ranged across both academia and industry, with 31 teams (69%) comprised of academic members only, 10 teams (22%) consisting of industrial members only, and 4 teams (9%) containing a mix of both. Teams were relatively small with all but one team (98%) having at most 6 members and 28 teams (62%) consisting of 3 or fewer members. Finally, in terms of effort, 20 teams (44%) reported spending less than 1 month developing their solutions, with 13 (29%) and 12 (27%) teams devoting 1-2 months and 3-6 months, respectively. The MTRAGEval task was well-received with 4.5/5 stars on the benchmark and 4.8/5 stars on the importance of the task.

5 Evaluation

We provided the teams with an evaluation script to run on their own on the development set. During the evaluation period, we provided the teams with the test data without the reference answers. Due to the sequential nature of the subtasks, we had two phases, a retrieval phase (Subtask A, C), followed by a generation phase (Subtask B). We only evaluated one submission per subtask, but participants could submit up to 3 times per subtask until the end of the evaluation phase and we only evaluated their final submission. Once the evaluation phase was complete, we evaluated each subtask on the same script provided to the participants with the reference passages and answers. We shared with each participating team their leaderboard score and rank as well as the score of the best baseline system and the reference files and their scored files for further analysis and evaluation for their paper submission.

The retrieval Subtask A was evaluated using the

common retrieval metrics, nDCG and Recall, with ranking based on nDCG@5. The two generation subtasks B and C were evaluated using the three main metrics reported in the MTRAG paper (Katsis et al., 2025), (1) \mathbf{RB}_{alg} : The harmonic mean of Bert-Recall, Rouge_L, and Bert-K-Prec, (2) \mathbf{RB}_{llm} : A Reference-Based LLM judge adapted from RAD-Bench (Kuo et al., 2025) and (3) \mathbf{RL}_F : The RAGAS (Es et al., 2024) Faithfulness LLM judge. All metrics are conditioned on answerability as shown in Table 3, using conditioning mechanisms from (Katsis et al., 2025; Rosenthal et al., 2026). For answerable, partials, and unanswerable questions, we use an IDK LLM judge that determines if the response contains an answer. For underspecified questions, we use a separate Clarification Judge to evaluate whether the response explicitly requests missing information to resolve the ambiguity.

	Condition	Score
A = answerable, partial	IDK Judge = yes	0
	IDK Judge = no, partial	ϕ
A = unanswerable	IDK Judge = yes	1
	IDK Judge = no, partial	0
A = underspecified	Clarification Judge = yes	1
	Clarification Judge = no	0

Table 3: Scoring mechanism for conditioning of metrics by answerability type. ϕ denotes the metric score (\mathbf{RB}_{alg} , \mathbf{RB}_{llm} , or \mathbf{RL}_f).

Note that the underspecified questions were excluded from the leaderboard and the rank was computed using the harmonic mean of the three metrics. In addition, due to the reference-less nature of the full RAG subtask we performed a small human evaluation (on 20 conversational tasks) for all participating teams for Subtask C (29). Full evaluation details are available in the benchmark paper (Katsis et al., 2025; Rosenthal et al., 2026), with implementation details including the evaluation pipeline code and LLM judge prompts publicly available in our repository <https://github.com/IBM/mt-rag-benchmark>.

6 Results

Table 4, 5, and 6 present the complete leaderboards with rank and scores for Subtasks A, B, and C. We also report several retrieval and generation baselines for the respective subtasks. Information regarding our selection of baselines can be found in MTRAG-UN (Rosenthal et al., 2026). We highlight individual top performing submissions in Section 8.

#	Team	nDCG@5	Recall@5	Approach
1	AILS-NTUA	0.578	0.622	Sparse
2	uva-irlab-conv	0.548	0.603	Sparse
3	Sifei	0.545	0.599	Hybrid
4	IIMAS-RAG	0.545	0.591	Hybrid
5	CuriosAI	0.540	0.586	Hybrid
6	GUIR	0.539	0.589	Sparse
7	TechSSN	0.534	0.582	Hybrid
8	caraman	0.531	0.559	Hybrid
9	BL.Research	0.505	0.552	Hybrid
10	GenAlus	0.496	0.548	Hybrid
11	ragtum	0.488	0.532	Dense
12	UTRAG	0.486	0.518	Dense
–	<i>Elser</i>	<i>0.480</i>	<i>0.521</i>	–
13	GigitAI	0.478	0.512	Hybrid
14	5ting	0.472	0.502	Dense
15	king001	0.470	0.516	Dense
16	PingAn	0.466	0.507	Dense
17	NLP-CEIA-UFG	0.450	0.489	Hybrid
18	HexU	0.449	0.481	Hybrid
19	uir_cis	0.442	0.490	Dense
20	PFW Task 8	0.433	0.474	Hybrid
21	H-RAG	0.427	0.479	Hybrid
22	HowardU-AI4PC	0.410	0.445	Hybrid
23	Polito Team	0.405	0.462	Dense
24	IITKanBDone	0.402	0.442	Sparse
–	<i>Granite R2</i>	<i>0.398</i>	<i>0.438</i>	–
–	<i>BGE-base 1.5</i>	<i>0.390</i>	<i>0.425</i>	–
25	SlugRAG	0.385	0.403	Dense
26	TDSA_UIT	0.379	0.409	–
27	mt-embed	0.368	0.395	Dense
–	<i>BM25</i>	<i>0.354</i>	<i>0.388</i>	–
28	evalx	0.347	0.391	–
29	YNU-HPCC	0.341	0.367	Hybrid
30	TTI-KDE	0.341	0.373	Dense
31	clulab-retrieval	0.331	0.360	Dense
32	RAID	0.285	0.312	Hybrid
33	RAGonauts	0.284	0.321	Lexical
34	445_grads	0.229	0.258	Sparse
35	JCT	0.175	0.217	Hybrid
36	kkanbu	0.137	0.171	–
37	Stratium	0.024	0.029	Dense
38	AKCIT - UFG	0.0	0.0	Dense

Table 4: Subtask A Leaderboard. Ranking is according to nDCG@5. Italicised rows are baselines.

6.1 Subtask A

A total of 38 teams participated in Subtask A with submissions ranked by nDCG@5 in Table 4. AILS-NTUA achieved the highest score with an nDCG@5 of 0.578 and Recall@5 of 0.622, outperforming the best baseline with Elser (0.480 nDCG@5 and 0.521 Recall@5) by a notable margin. uva-irlab-conv and Sifei followed closely at 0.548 and 0.545 respectively, with the top 12 teams all exceeding the best baseline. Some teams such as AKCIT-UFG, Stratium, and kkanbu received zero or low scores as almost all of their retrieved passages did not align with the evaluation reference passages. For instance, AKCIT-UFG introduced a custom chunking strategy and modified the origi-

nal document IDs, while kkanbu used a different collection than what the query was intended for.

Hybrid retrieval was the most popular approach, used by 18 of 38 teams (including variants with explicit reranking). Dense retrieval was the second most common, used by 13 teams (including one team that employed HyDE-based dense search with FAISS). Sparse retrieval was used by 5 teams, while 1 team relied on pure lexical (BM25-only) retrieval. Two of the top three teams, AILS-NTUA and uva-irlab-conv used sparse retrieval rather than dense or hybrid methods. This aligns with our baseline experiments that show Sparse methods outperform dense methods. As described in MTRAG (Katsis et al., 2025), a sparse retriever was used during benchmark creation which may give preference to passages retrieved from sparse approaches. The consistent theme across top performers is the use of query rewriting to resolve coreferences and produce self-contained queries prior to retrieval, reinforcing findings from prior work on multi-turn retrieval (Katsis et al., 2025).

Figures 4 and 5 in the Appendix show detailed results per domain and standalone/non-standalone. The Subtask A domain heatmap reveals both a clear overall hierarchy and meaningful cross-domain specialization. AILS-NTUA leads overall and dominates in all domains. FiQA is uniformly the hardest domain. Domain rank reversals are pronounced and informative: TechSSN ranks 14th on ClapNQ; uva-irlab-conv falls to 13th on FiQA (0.330); GigitAI ranks 21st in ClapNQ (0.470). These divergences suggest that collection-specific characteristics can drive meaningfully different behaviors across retrieval systems. The coref-sensitivity chart shows that the majority of Subtask A systems are hurt by coreference-dependent queries. Five teams exhibit a meaningful positive delta suggesting these systems effectively exploit conversational history, perhaps through explicit query rewriting or context-aware representations. Interestingly, ragtum’s pattern on standalone questions (+0.058) suggest overly aggressive query rewriting - knowing when to not rewrite is also important. The remaining majority (14 teams) are harmed by non-standalone queries, with the largest degradations among lower-ranked teams.

6.2 Subtask B

Subtask B evaluated generation quality given gold retrieved passages. The 26 participating teams were ranked by the harmonic mean (H-Mean) of

#	Team	Harm Mean	RB alg	RL F	RB llm	Model Family	Gen
1	RaguTeam	0.783	0.64	0.93	0.84	Many	Many
2	AILS-NTUA	0.770	0.63	0.90	0.83	GPT	ZS
3	YNU-HPCC	0.768	0.64	0.90	0.82	DS	FS
4	GenAIus	0.763	0.61	0.89	0.86	GPT	ZS
5	CuriosAI	0.757	0.59	0.91	0.86	GPT	FS
6	PFW Task 8	0.756	0.61	0.89	0.83	GPT	ZS
7	GigitAI	0.750	0.61	0.89	0.80	Many	ZS
8	Mind_Flayer	0.749	0.60	0.88	0.83	Llama	ZS
9	HowardU-AI4PC	0.746	0.63	0.85	0.79	Llama	ZS
10	RAID	0.745	0.60	0.86	0.84	DS	ZS
11	ragtum	0.745	0.59	0.90	0.82	Gemini	ZS
12	king001	0.730	0.57	0.88	0.83	Gemini	ZS
13	IITKanBDone	0.698	0.58	0.79	0.78	Mistral	FS
14	Team V	0.659	0.50	0.78	0.79	Qwen	ZS
15	UTRAG	0.655	0.55	0.78	0.69	Qwen	SF
-	<i>GPT-OSS-120B</i>	<i>0.639</i>	<i>0.49</i>	<i>0.70</i>	<i>0.82</i>	GPT	ZS
16	Polito Team	0.638	0.48	0.80	0.73	Gemma	ZS
17	Dawn	0.633	0.49	0.75	0.74	GPT	ZS
18	TeamRagnition	0.618	0.49	0.70	0.72	Qwen	ZS
-	<i>Mistral-Small-4B</i>	<i>0.604</i>	<i>0.48</i>	<i>0.68</i>	<i>0.72</i>	Mistral	ZS
19	NCL	0.597	0.50	0.72	0.61	Qwen	ZS
-	<i>GPT-OSS-20B</i>	<i>0.590</i>	<i>0.46</i>	<i>0.65</i>	<i>0.73</i>	GPT	ZS
-	<i>Granite-4-Small</i>	<i>0.583</i>	<i>0.49</i>	<i>0.69</i>	<i>0.60</i>	Granite	ZS
-	<i>DeepSeek-V3</i>	<i>0.581</i>	<i>0.46</i>	<i>0.68</i>	<i>0.67</i>	DS	ZS
20	HexU	0.571	0.46	0.68	0.63	Qwen	FS
-	<i>Qwen-30B-Think</i>	<i>0.571</i>	<i>0.43</i>	<i>0.65</i>	<i>0.73</i>	Qwen	ZS
-	<i>Llama-3.3-70B</i>	<i>0.567</i>	<i>0.46</i>	<i>0.66</i>	<i>0.62</i>	Llama	ZS
21	uir_cis	0.566	0.45	0.65	0.65	Qwen	ZS
22	IIMAS-RAG	0.564	0.43	0.69	0.64	GPT	ZS
-	<i>Qwen3-8B</i>	<i>0.540</i>	<i>0.42</i>	<i>0.60</i>	<i>0.67</i>	Qwen	ZS
-	<i>Phi-4</i>	<i>0.534</i>	<i>0.42</i>	<i>0.57</i>	<i>0.68</i>	Phi	ZS
-	<i>Mistral-Large-675B</i>	<i>0.523</i>	<i>0.40</i>	<i>0.56</i>	<i>0.70</i>	Mistral	ZS
-	<i>Llama-4-17B</i>	<i>0.515</i>	<i>0.40</i>	<i>0.62</i>	<i>0.58</i>	Llama	ZS
23	uva-irlab-conv	0.512	0.37	0.60	0.68	Other	ZS
-	<i>Granite-4-Tiny</i>	<i>0.476</i>	<i>0.39</i>	<i>0.53</i>	<i>0.54</i>	Granite	ZS
24	Stratium	0.442	0.37	0.52	0.46	Many	ZS
25	BL.Research	0.427	0.39	0.36	0.60	GPT	FS
-	<i>DeepSeek-R1</i>	<i>0.372</i>	<i>0.26</i>	<i>0.40</i>	<i>0.57</i>	DS	ZS
26	VerbanexAI	0.226	0.23	0.23	0.23	DS	ZS

Table 5: Subtask B Leaderboard. Harmonic Mean ranking was computed using four decimal points. DS = DeepSeek, Generation Method (Gen): ZS = Zero-Shot, FS = Few-Shot, SF = Supervised Fine-Tuning, RL = Reinforcement Learning. Italicised rows are baselines.

three metrics: RB_{alg} , RL_F , and RB_{llm} (Table 5). RaguTeam achieved the highest H-Mean of 0.783, followed by AILS-NTUA (0.770) and YNU-HPCC (0.768). The top 15 teams all exceeded the strongest LLM baseline, GPT-OSS-120B (0.639), demonstrating that many participating teams substantially outperformed off-the-shelf generation approaches (Rosenthal et al., 2026).

Among the 26 participating teams, open-source models were predominant (15 teams), with the remaining 11 relying on proprietary or commercially available APIs. Teams used a variety of models for generation with the most popular model families being GPT and Qwen, each adopted by 8 (31%)

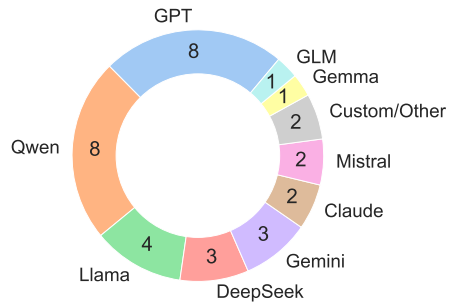


Figure 2: Model families used for generation by teams in Subtask B (some teams used multiple models).

teams (see Figure 2). A few (3) teams used several models for generation, with the top-ranked team employing an ensemble of 7 different LLMs. Zero-shot prompting was by far the most common generation strategy, adopted by 19 teams, while 5 teams used few-shot prompting and 1 used supervised fine-tuning. Finally, the top-ranked team used an ensemble of all three strategies.

In terms of interesting patterns, 3 out of the top 5 teams created multiple candidate responses and used LLM judge and/or other metrics to select the final answer. Moreover, several teams employed an agentic approach, splitting the generation process into multiple stages. Common stages included determining answerability of the user question as well as rewriting it.

Figures 6 and 7 in the Appendix show detailed results per answerability and domain. The Subtask B answerability breakdown reveals a clear and consistent hierarchy: ANSWERABLE queries yield the highest scores, followed by PARTIAL and UNANSWERABLE. Performance varies more across teams on UNANSWERABLE queries than ANSWERABLE; top systems maintain competitive unanswerable scores, while many mid-tier teams drop sharply, suggesting that abstention is a differentiating capability not correlated with overall generation quality. The PARTIAL category also shows notable spread, with some teams generating strong partial answers and others exhibiting a sharp drop from their ANSWERABLE performance, likely reflecting differences in how systems handle incomplete evidence. Fig 7 shows that ClapNQ and Govt tend to yield higher scores than FiQA and IBMCloud, suggesting that generation challenges differ depending on the domain.

6.3 Subtask C

Subtask C presented the most challenging setting, requiring systems to perform both retrieval

#	Team	Harm Mean	RB alg	RL F	RB llm	Model Family	Gen
1	GenAIus	0.586	0.43	0.74	0.68	GPT	ZS
2	HowardU-AI4PC	0.580	0.44	0.75	0.63	Llama	ZS
3	5ting	0.560	0.39	0.77	0.68	GPT	ZS
4	DUTIR	0.558	0.41	0.71	0.66	GPT	ZS
5	GigitAI	0.556	0.44	0.70	0.60	Claude	ZS
6	TTI-KDE	0.552	0.42	0.75	0.59	Gemini	ZS
7	CuriosAI	0.549	0.38	0.76	0.67	GPT	FS
8	king001	0.548	0.39	0.68	0.69	Gemini	ZS
9	PingAn	0.543	0.39	0.66	0.69	Gemini	SF
10	ragtum	0.542	0.38	0.74	0.64	Gemini	ZS
11	AILS-NTUA	0.541	0.40	0.73	0.60	GPT	ZS
12	GUIR	0.540	0.41	0.68	0.61	Claude	ZS
13	IIMAS-RAG	0.540	0.40	0.71	0.61	GPT	ZS
–	<i>Qwen-30B-Think</i>	<i>0.537</i>	<i>0.39</i>	<i>0.67</i>	<i>0.66</i>	Qwen	ZS
–	<i>Mistral-Small-4B</i>	<i>0.534</i>	<i>0.40</i>	<i>0.65</i>	<i>0.63</i>	Mistral	ZS
14	PFW Task 8	0.533	0.42	0.67	0.57	GPT	ZS
15	Sifei	0.531	0.38	0.73	0.60	DS	ZS
–	<i>GPT-OSS-120B</i>	<i>0.527</i>	<i>0.38</i>	<i>0.63</i>	<i>0.69</i>	GPT	ZS
–	<i>Llama-3.3-70B</i>	<i>0.524</i>	<i>0.40</i>	<i>0.64</i>	<i>0.60</i>	Llama	ZS
–	<i>DeepSeek-V3</i>	<i>0.522</i>	<i>0.38</i>	<i>0.64</i>	<i>0.63</i>	DS	ZS
–	<i>Granite-4-Small</i>	<i>0.516</i>	<i>0.41</i>	<i>0.60</i>	<i>0.58</i>	Granite	ZS
16	UTRAG	0.516	0.40	0.70	0.53	Qwen	SF
–	<i>GPT-OSS-20B</i>	<i>0.514</i>	<i>0.38</i>	<i>0.59</i>	<i>0.67</i>	GPT	ZS
17	HexU	0.506	0.39	0.64	0.57	Qwen	FS
18	evalx	0.505	0.38	0.65	0.56	–	–
19	RAID	0.504	0.36	0.68	0.58	DS	ZS
–	<i>Qwen3-8B</i>	<i>0.498</i>	<i>0.37</i>	<i>0.58</i>	<i>0.62</i>	Qwen	ZS
–	<i>Llama-4-17B</i>	<i>0.492</i>	<i>0.37</i>	<i>0.60</i>	<i>0.59</i>	Llama	ZS
20	uva-irlab-conv	0.487	0.32	0.66	0.65	Other	ZS
21	Polito Team	0.485	0.33	0.62	0.65	Gemma	ZS
–	<i>Phi-4</i>	<i>0.479</i>	<i>0.36</i>	<i>0.54</i>	<i>0.62</i>	Phi	ZS
–	<i>Mistral-Large-675B</i>	<i>0.473</i>	<i>0.34</i>	<i>0.54</i>	<i>0.64</i>	Mistral	ZS
22	Model Citizens	0.455	0.37	0.57	0.47	GPT	ZS
–	<i>Granite-4-Tiny</i>	<i>0.449</i>	<i>0.34</i>	<i>0.52</i>	<i>0.55</i>	Granite	ZS
23	IITKanBDone	0.429	0.31	0.61	0.47	Mistral	FS
24	NLP-CEIA-UFG	0.377	0.27	0.47	0.47	Many	ZS
–	<i>DeepSeek-RL</i>	<i>0.351</i>	<i>0.23</i>	<i>0.40</i>	<i>0.56</i>	DS	ZS
25	H-RAG	0.324	0.25	0.27	0.65	GPT	ZS
26	TechSSN	0.320	0.24	0.60	0.28	–	–
27	MingRAG	0.319	0.29	0.38	0.31	Llama	RL
28	BL.Research	0.292	0.25	0.24	0.47	GPT	FS
29	Stratium	0.198	0.22	0.19	0.19	Mistral	ZS

Table 6: Subtask C Leaderboard. Harmonic Mean ranking was computed using four decimal points. DS = DeepSeek, Generation Method (Gen): ZS = Zero-Shot, FS = Few-Shot, SF = Supervised Fine-Tuning, RL = Reinforcement Learning. Italicised rows are baselines. Teams with – did not complete the registration survey.

and generation end-to-end, which is reflected in the overall lower scores compared to Subtask B. Among 29 participating teams (Table 6), GenAIus achieved the highest H-Mean of 0.586, followed by HowardU-AI4PC (0.580) and 5ting (0.560). In contrast to Subtask B where open-source models were predominant, Subtask C saw a shift toward proprietary APIs, with 19 teams using non-open-source models. Hybrid retrieval remained the most common retrieval strategy, used by the majority of teams, followed by dense and sparse methods,

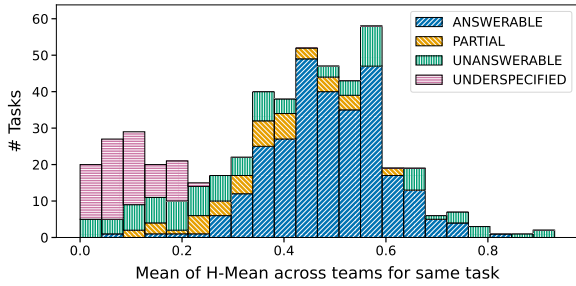
mirroring Subtask A trends. Zero-shot prompting was again overwhelmingly dominant (20 teams), with 4 teams using few-shot, 2 using supervised fine-tuning, and, notably, MingRAG using RL.

Figures 8 and 9 in the Appendix show detailed results per answerability and domain. GenAIus emerges as the Subtask C leader despite ranking 8th in Subtask A, suggesting its architecture is particularly well suited to multi-turn coherence and grounding. HowardU-AI4PC, ranked 19th in retrieval, also climbs dramatically. UNANSWERABLE performance diverges more sharply across teams in Subtask C than in Subtask B: the top Subtask C systems (GenAIus, HowardU-AI4PC) maintain competitive unanswerable-query scores, while many lower-ranked teams fall to near-zero, suggesting that correctly recognizing unanswerable questions becomes harder—and more discriminating—when conversational context is present. ANSWERABLE scores remain the strongest column but are generally lower in absolute terms than Subtask B equivalents. The PARTIAL category is the most tightly distributed in Subtask C, as fewer teams achieve strong partial-answer performance than in Subtask B. Domain patterns mirror Subtask B but with increased variance, and some domain-specific specializations seen in Subtask A and B are amplified. We provide insight into variations of team outputs for a single example in the Appendix, Table 9.

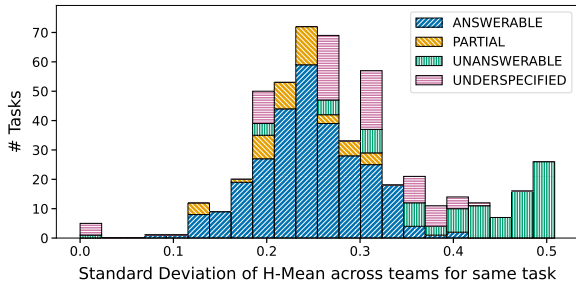
6.3.1 Score Distribution Analysis

To better understand the distribution of scores, we compute the mean and the standard deviation of the (H-Mean) scores for each team per conversational task. Figures 3a and 3b depict the distributions of the score mean and standard deviation of scores, respectively. For the purposes of this analysis, we used the entire test set, including Underspecified.

The score mean distribution shows that the hardest conversational tasks for most systems (lowest mean score) were underspecified and unanswerable questions. Moreover, the distribution of standard deviation shows that systems in general had low-to-medium disagreement (and thus standard deviation) in their scores, with the exception of unanswerable questions, showing higher variation. This is likely due to some models getting the answerability of unanswerable questions wrong (i.e., producing non-IDK answers to unanswerable questions) and therefore receiving a zero score by the IDK-conditioning used in our metrics, while other mod-



(a) Mean of H-Mean scores.



(b) Standard deviation of H-Mean scores.

Figure 3: Distribution of mean/standard deviation of H-Mean scores for Subtask C across teams by answerability category.

els treated them correctly as unanswerables, producing an IDK response and receiving full marks.

6.3.2 Subtask C Human Evaluation

One challenge in evaluating Subtask C is that, without a reference answer, each system can retrieve different passages, complicating direct comparisons. To assess response quality in this referenceless setting, we performed a human evaluation on a subsample of 20 conversational tasks, with an emphasis on tasks that had poor and mixed performance, for 28 models and the reference. It is important to note that since there are a small amount of questions, this should be looked at only as an additional assessment on system quality. The algorithmic metrics reported over all the conversational tasks in Table 6 are considerably more stable, as individual questions have negligible impact, and is therefore used as the official ranking for the task.

The annotation task was done on the Appen crowdsourcing platform, where each system response was independently assessed by three annotators. The eight annotators that contributed to this work are highly skilled individuals hired solely to perform language annotation and paid well above minimum wage. Annotators were presented with the multi-turn conversation context, the retrieved passages, and the model’s response. The annotators

#	Team	Overall	FANC	#	HM
1	Reference Gold	3.42 ± 0.32	3.65 ± 0.23	1	0.82
2	IIMAS-RAG	3.33 ± 0.50	3.80 ± 0.16	3	0.65
3	king001	3.10 ± 0.42	3.61 ± 0.25	9	0.55
4	5ting	3.06 ± 0.29	3.67 ± 0.16	18	0.48
5	HowardU-AI4PC	3.04 ± 0.44	3.65 ± 0.27	2	0.66
6	TTI-KDE	3.04 ± 0.34	3.43 ± 0.30	5	0.60
7	GUIR	3.02 ± 0.42	3.41 ± 0.23	5	0.60
8	DUTIR	2.96 ± 0.38	3.61 ± 0.26	8	0.57
9	ragtum	2.96 ± 0.37	3.46 ± 0.33	9	0.55
10	AILS-NTUA	2.94 ± 0.47	3.31 ± 0.40	18	0.48
11	GigitAI	2.92 ± 0.23	3.61 ± 0.23	4	0.61
12	uva-irlab-conv	2.90 ± 0.56	3.33 ± 0.25	21	0.43
13	CuriosAI	2.90 ± 0.51	3.59 ± 0.28	21	0.43
14	PingAn	2.90 ± 0.34	3.52 ± 0.31	9	0.55
15	UTRAG	2.88 ± 0.40	3.45 ± 0.25	9	0.55
16	GenAIus	2.79 ± 0.37	3.56 ± 0.24	20	0.46
17	Sifei	2.77 ± 0.46	3.42 ± 0.36	7	0.58
18	HexU	2.73 ± 0.48	3.55 ± 0.25	17	0.50
19	evalx	2.71 ± 0.25	3.46 ± 0.17	24	0.39
20	RAID	2.67 ± 0.66	3.37 ± 0.34	24	0.39
21	IITKanBDone	2.60 ± 0.70	3.02 ± 0.52	13	0.52
22	PFW Task 8	2.58 ± 0.48	3.46 ± 0.36	15	0.51
23	Polito Team	2.56 ± 0.39	3.35 ± 0.27	26	0.37
24	Model Citizens	2.54 ± 0.32	3.67 ± 0.34	13	0.52
25	H-RAG	2.42 ± 0.64	2.59 ± 0.35	27	0.33
26	NLP-CEIA-UFG	2.40 ± 0.45	3.03 ± 0.50	15	0.51
27	BL.Research	2.19 ± 0.47	2.26 ± 0.33	28	0.31
28	TechSSN	2.04 ± 0.21	3.09 ± 0.46	23	0.41
29	Stratium	1.75 ± 0.64	2.17 ± 0.50	29	0.23

Table 7: Subtask C results: human evaluation (IDK conditioned) on 16 examples with Subtask C Harmonic Mean (HM) on the subset for comparison.

first selected which retrieved passages, if any, were relevant to answering the question given the conversation and also indicated whether the response indicated IDK (I don’t know). Then they were asked to rate the response on four dimensions using a 4-point scale (1 = No, 2 = Mostly No, 3 = Mostly Yes, 4 = Yes): Faithfulness (whether the response is grounded in the retrieved passages), Appropriateness (whether the response provides an appropriate amount of useful information), Naturalness (whether the response is coherent and not dismissive), and Completeness (whether the response includes all relevant information found in the context) (Katsis et al., 2025). Annotators also provided an Overall Rating on the same 4-point scale and indicated whether the model declined to answer by saying it did not know. The annotators did not know what the reference answer was, instead it was provided as an additional model.

Annotator relevance judgments were analyzed across all 20 tasks to assess alignment between retrieved passages and task answerability. For answerable tasks, most systems correctly retrieved relevant passages. Analysis revealed one answer-

able task with no relevant passages, even for the reference response, and one unanswerable task where 27 models found relevant passages. We also computed IDK rates, revealing how well models are calibrated to answerability. Most models show the expected directional pattern: On unanswerable tasks, 15/28 models correctly abstained on most tasks, while 13 abstained on half or fewer. On answerable tasks, most models provided an answer, though 5 models said IDK on more than 44% of answerable tasks, indicating over-hedging. Partial tasks show the most variation across models, consistent with their ambiguous nature, and most models attempt to answer underspecified tasks rather than acknowledging the lack of sufficient information.

Table 7 presents the ranking¹ of the 28 participating systems and the Reference Gold baseline, on the human annotations tasks across two human evaluation metrics: Overall Rating and FANC (the harmonic mean of Faithfulness, Appropriateness, Naturalness, and Completeness). Scores are reported as mean \pm standard deviation on a 1–4 scale averaged across annotators and tasks. Reference Gold has the highest overall rating and comparable ranking for FANC to top-performing systems, suggesting that the best systems approach human-authored response quality. There is a notable spread across the leaderboard indicating substantial variation in response quality across systems.

6.4 Cross-Subtask Comparison

Comparing across subtasks, teams that participated in both Subtask B and Subtask C consistently scored lower in Subtask C, with an average H-Mean drop of approximately 0.20 points. For example, AILS-NTUA dropped from 0.770 to 0.541 (Ranks A: 1, B: 2, C: 11). This gap reflects the compounding effect of retrieval errors on generation quality in a multi-turn setting. However, doing well on retrieval doesn't impact full RAG as much as doing well on generation as evident by GenAIus (Ranks A: 11, B: 4, C: 1) performing best on full RAG but lower on retrieval. On the other hand AILS performed best on retrieval but lower on full RAG - however the difference between the scores is not as large as for Subtask A. TechSSN (Ranks A: 7, C: 26) performed well on Subtask A, but suffered for Subtask C because they answered IDK too often

¹Rankings are computed over all annotated tasks except for Underspecified for which the notion of correct response was not clearly defined to the annotators, and the two tasks with incorrect reference annotations.

(see Table 8).

7 Underspecified Questions

We included a surprise answerability type of 78 questions categorized as Underspecified, meaning the response to the question should be asking for further clarification i.e., explicitly request the missing information needed to resolve the ambiguity. We did not include this in the official leaderboard, but we report next general performance on this subset. The underspecified questions are evaluated by a dedicated judge determining whether the generated response asks for clarification. Responses that directly answer an underspecified question or expresses inability to answer without soliciting specifics, receive no credit (Rosenthal et al., 2026).

Out of 26 Subtask B teams, 13 scored zero on underspecified questions, including top-ranked teams such as RaguTeam, AILS-NTUA, and GenAIus. These teams generated direct answers in all cases rather than asking for clarification. Conversely, teams that performed well on underspecified questions tended to rank near the bottom overall, as they over-requested clarification even for well-specified questions where direct answer is expected. For instance, teams such as BL.Research (Subtask B, C) and H-RAG (Subtask C) performed well on the underspecified questions, while performing poorly on the overall leaderboard (Figures 7, 9 in Appendix).

These findings highlight the importance of evaluating underspecified questions as a separate dimension. A model that optimizes only for the overall generation quality may never learn to seek clarification. The consistent failure of most teams on underspecified questions underscores that knowing when to seek clarification rather than generating an answer remains an open and largely unaddressed challenge in multi-turn RAG.

8 Description of Selected Teams

In this section, we highlight the details of specific approaches for selected teams per subtask.

8.1 Subtask A

AILS-NTUA Ranked 1st, using a Sparse approach. Pipeline: 1) 5 different rewritings of the original query based on the conversation's history, 2) retrieval using Elser-v1 with each one rewrite, 3) rerank the top-100 results of each retrieval independently using cohere-rerank-v4.0-pro, 4) fusion of the results of the 5 different reranked lists using

weighted RRF and to get the top 10 documents. Other approaches tried including hybrid retrieval or LLM reranking hurt the main metrics.

Sifei Ranked 3rd, using a Hybrid approach. Applied data cleaning by removing non-English passages, near-duplicates, ultra-short fragments, with strict qrels protection. They evaluated 6 dense models and picked Jasper-600M. They evaluated BM25, SPLADE-v2, and SPLADE-v3, and picked SPLADE-v3. They tuned several parameters and tested multiple rerankers and also applied query rewriting and different fusion strategies.

UTRAG Ranked 12th, using a Dense approach. They used a two-stage query rewriting pipeline followed by dense retrieval and reranking. The rewriting included resolving coreference only when needed, and standalone rewrite uses coarse prompts for high recall and fine-grained prompts for better ranking. Dense retrieval was performed with FAISS retrieving top-50 candidates, with reranking to top-10 to balance Recall and nDCG.

8.2 Subtask B

RaguTeam Ranked 1st, using a heterogeneous ensemble of 7 LLMs, including Meno-Lite (a domain-adapted Qwen2.5-7B fine-tuned for RAG), Gemini-3-Pro-Preview, Claude 4.5 Haiku, GLM-4.6, Llama-3.3-70B, Qwen3-235B-A22B, and Qwen2.5-32B. They used GPT-4o-mini as a judge to select the best response per instance.

AILS-NTUA Ranked 2nd, using an agentic approach with a 6-stage pipeline: determining answerability, extracting spans with DeepSeek-V3.2, generating dual answers with GPT-4o, evaluating with both a technical judge (DeepSeek) and a human-like judge (GPT-4o-mini), selecting the best answer, and applying final refinements via GPT-4o-mini.

YNU-HPCC Ranked 3rd, using BGE-M3-based dynamic few-shot learning with ROUGE-guided self-consistency to improve extraction fidelity, powered by DeepSeek-V3.

8.3 Subtask C

GenAIus Ranked 1st, using hybrid retrieval and zero-shot prompting. The user query was used for multiple rewrites for lexical and semantic search. Retrieval combined BM25, HyDE generation, and dual semantic retrieval. Semantic search using

Qwen3-Embedding-0.6B and text-embedding-3-small was performed over the top lexical candidates, fused via Reciprocal Rank Fusion, and re-ranked using an LLM relevance judge. Generation used GPT-4o, conditioned on the conversation history, the last user question, its semantic rewriting, and the retrieved context.

Howard University-AI4PC Ranked 2nd, using hybrid retrieval and zero-shot prompting. They retrieved using Dense BGE embeddings and BM25 lexical search fused via Reciprocal Rank Fusion, HyDE, query decomposition for complex multi-part questions, LLM-based query rewriting for multi-turn conversations, and cross-encoder reranking with BGE-reranker-large. Generation was done by zero-shot prompting of Llama-3.3-70B-Instruct, grounding answers strictly in retrieved passages.

CuriosAI Ranked 7th, using hybrid retrieval and few-shot (FS) prompting. The corpus was preprocessed, summarized, and queries were rewritten using GPT-5. Four embedding models were used to create four vector indices over the summary-appended corpus. Hybrid Retrieval (RRF) combined the top-100 results from the models for the top-10 results. For generation, GPT-5 was prompted using a FS prompt and sampling five candidate answers. The final answer was selected by maximizing the harmonic mean of RAGAS faithfulness and LLM-as-a-judge scores.

9 Conclusion

MTRAGEval was a successful SemEval task attracting strong participation with 107 registered teams and 92 submissions for all subtasks across 48 teams. In Subtask A, reranking and query rewriting were important. The best approaches for Subtask B and C used an ensemble of models without training for the subtask. All methods struggled with the surprise underspecified queries, and unanswerable questions continue to be a challenge. There is still room for improvement across all subtasks. Our benchmark is publicly available to the community at <https://github.com/IBM/mt-rag-benchmark>.

References

Pasha Abrishamchian, Ophir Frieder, and Nazli Goharian. 2026. GUIR at semeval-2026 task 8: Training-free multi-query fusion for robust conversational retrieval. In *Proceedings of the 20th International*

- Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. **TREC iKAT 2023: A test collection for evaluating conversational and interactive knowledge assistants**. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, page 819–829, New York, NY, USA. Association for Computing Machinery.
- Dimosthenis Athanasiou, Maria Lymperaioi, George Filandrianos, Athanasios Voulodimos, and Giorgos Stamou. 2026. AILS-NTUA at semeval-2026 task 8: Query diversity via nested RRF and evidence-guided agentic generation for multi-turn RAG. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- David Maximilian Caraman and Gheorghe Cosmin Silaghi. 2026. Caraman at semeval-2026 task 8: Three-stage multi-turn retrieval with query rewriting, hybrid search, and cross-encoder reranking. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Roman Derunets, Mikhail Komarov, Ivan Chernov, Oleg Sedukhin, Ivan Bondarenko, and Mikhail Kulakov. 2026. RaguTeam at semeval-2026 task 8: Meno and friends in a judge-orchestrated LLM ensemble for faithful multi-turn response generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Guilherme C. Dutra, André Felipe dos Santos Caraíba, Nádia Félix F. da Silva, Paulo V. dos Santos, Deborah Silva A. Fernandes, and Sávio Salvarino T. de Oliveira. 2026. NLP-CEIA-UFG at semeval-2026 task 8: Iterative retrieval exposes critical overabstention in multi-turn RAG. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. **FaithDial: A faithful benchmark for information-seeking dialogue**. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.
- Passant Elchafei, Hossam Emam, Mohamed Alansary, Monorama Swain, and Markus Schedl. 2026. H-RAG at semeval-2026 task 8: Hierarchical parent-child retrieval for multi-turn RAG conversations. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. **RAGAs: Automated evaluation of retrieval augmented generation**. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.
- Kshitij Fadnis, Sara Rosenthal, Maeda Hanafi, Yannis Katsis, and Marina Danilevsky. 2025. **Ragaphene: A rag annotation platform with human enhancements and edits**. *Preprint*, arXiv:2508.19272.
- Tal Farhan and Chaya Liebeskind. 2026. JCT at semeval-2026 task 8: Resource-efficient multi-turn RAG via nano-LLM rewriting and hybrid reranking. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. **MultiDoc2Dial: Modeling dialogues grounded in multiple documents**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- David Oneil Ferreira, Wilson Maranhão Ramos, Priscila Maia Ribeiro, Emanuel Borges Passinato, diogo fernandes, and Arlindo Rodrigues Galvão Filho. 2026. AKCIT-UFG at semeval-2026 task 8: Structured chunking and optimized query reformulation for efficient multi-turn retrieval. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Anne Jacika J, Anishka K, Guruprakash K, Rajalakshmi Sivanaiah, and Angel Deborah S. 2026. TechSSN at semeval-2026 task 8: MTRAG retrieval and generation using ensemble re-encoders and anchor prompting. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. **MTRAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems**. *Transactions of the Association for Computational Linguistics*, 13:784–808.
- Hyungji Kim, Siva Rohit Kondapaneni, and Steven Bethard. 2026. clulab-retrieval at semeval-2026 task 8: A comparative analysis of dense retrievers and

- hyde for multi-turn conversational retrieval. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Saran Krishnasamy and Inez Wihardjo. 2026. GigitAI at semeval-2026 task 8: Hybrid sparse-dense retrieval and zero-shot generation for multi-turn conversational RAG. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Aiswariya Manoj Kumar, Hiroki Takushima, Fumika Beppu, Yuki Shibata, Daichi Yamaga, and Takayuki Hori. 2026. CuriousAI at semeval-2026 task 8: Hybrid retrieval system with repeated sampling for generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Tzu-Lin Kuo, FengTing Liao, Mu-Wei Hsieh, Fu-Chieh Chang, Po-Chun Hsu, and Da-shan Shiu. 2025. **RAD-bench: Evaluating large language models' capabilities in retrieval augmented dialogues**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 868–902, Albuquerque, New Mexico. Association for Computational Linguistics.
- Bo Li, You Zhang, Jin Wang, Dan Xu, and XueJie Zhang. 2026a. YNU-HPCC at semeval-2026 task 8: Parallel generation and multi-metric reranking for faithful extractive RAG. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Feiling Li, Xiaoya Qi, Xunyu Wang, Pusheng Chen, Zhiwen Tang, and Han Yang. 2026b. Dawn at semeval-2026 task 8: Structured control decomposition for faithful multi-turn retrieval-augmented generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Zehao Liu and Huizhi Liang. 2026. NCL at semeval-2026 task 8: Deterministic small-LLM RAG with relation classification. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Simon Lupart, Kidist Amde Mekonnen, Zahra Abbasiantaeb, and Mohammad Aliannejadi. 2026. uva-irlab-conv at semeval-2026 task 8: Multi-turn RAG with learned sparse retrieval and listwise reranking. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Sifei Meng and Dmitry Ilvovsky. 2026. Sifei at semeval-2026 task 8: Hybrid retrieval and query rewriting for multi-turn rag. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghadam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. **Ragnarök: A reusable rag framework and baselines for trec 2024 retrieval-augmented generation track**. *Preprint*, arXiv:2406.16828.
- Soumendra Nath Ray and Garima Gupta. 2026. IITKanBDone at semeval-2026 task 8: Elser-based sparse retrieval with mistral-7b for multi-turn RAG evaluation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Vania Raya-Rios, Helena Gomez-Adorno, Leon Hecht, Pedro Vázquez-Osorio, Erick Fabián-Sandoval, Jesús Vázquez-Osorio, and Diego Hernández-Bustamante. 2026. IIMAS-RAG at semeval-2026 task 8: Hybrid sparse-dense retrieval and answerability-conditioned generation for multi-turn RAG. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Pratibha Revankar, Jihye Kim, and Umit Azirakhmet. 2026. SlugRAG at semeval-2026 task 8: Domain-specific fine-tuning and model scaling for multi-turn RAG retrieval. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Sara Rosenthal, Yannis Katsis, Vraj Shah, Lihong He, Lucian Popa, and Marina Danilevsky. 2026. **MTRAG-UN: A benchmark for open challenges in multi-turn rag conversations**. *Preprint*, arXiv:2602.23184.
- Sijan Shrestha and Saurav K. Aryal. 2026. Howard University-AI4PC at semeval-2026 task 8: Query reformulation and dense-lexical retrieval fusion for multi-turn retrieval-augmented generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Rajalakshmi Sivanaiah, Angel Deborah S, Karthik Raja C, and Rithika S. 2026. RAGonauts at semeval-2026 task 8: Bm25 retrieval with last-turn query formulation for multi-turn RAG conversations. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Thien-Qua T-Nguyen, Chi Hoang, Nguyen Thi Bao Tran, Tri Le, Khanh Phuoc Bao Truong, and

- Chinh Trong Nguyen. 2026. 5ting at semeval-2026 task 8: Strong end-to-end multi-turn RAG via LLM-based reranking and faithfulness control. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Taleef Tamsal. 2026. PFW Task 8 at semeval-2026 task 8: Lightweight tri-fusion retrieval with prompt-engineered faithful generation for multi-turn RAG. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Jerin Romijah Tuli, MD. Sartaj Alam Pritom, and Talukder Naemul Hasan Naem. 2026. Mind_Flayer at semeval-2026 task 8:dualrag:answerability-aware generation for multi-turn RAG conversations. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Finn Wigger, Maximilian Podolsky, Merle Wilmlink, Zelong Peng, Shaghayegh Kolli, and Jana Diesner. 2026. RAGTUM at semeval-2026 task 8: Contextual query rewriting and dense retrieval for multi-turn RAG. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Jin Rui Yang, Yichong Chen, and Liang Yang. 2026. DUTIR at semeval-2026 task 8: A hybrid retrieval and faithfulness-guarded framework for multi-turn RAG. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Suveyda Yeniterzi and Reyhan Yeniterzi. 2026. GenAIus at semeval-2026 task 8: Beyond retrieval with relevance-aware RAG for faithful multi-turn generation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Jiaqi Zhang, Wenbin Duan, Yingqi Zhang, Yan Li, and Binyang Li. 2026. uir_cis at semeval-2026 task 8: A unified lightweight pipeline for multi-turn RAG evaluation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Ke Zhou and Yi-Shan Lin. 2026. UTRAG at semeval-2026 task 8: History-aware query rewriting and lora-finetuned generation for multi-turn rag. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.
- Murat Çelik, Nejla Dinçer, Can Ersoy, Mert Toprak, Barış Tan Ünal, Riccardo Coppola, and Flavio Giobergia. 2026. Polito Team at semeval-2026 task 8: Scaling multi-turn RAG: High-performance parallelized pipeline for the MTRAG benchmark. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California, United States. Association for Computational Linguistics.

Team (by overall rank)	Domain			
	Overall	ClapNQ	Govt	FiQA
AILS-NTUA	0.578 (1)	0.63 (3)	0.63 (1)	0.46 (1)
uva-irlab-conv	0.548 (2)	0.64 (2)	0.59 (3)	0.33 (1)
IMAS-RAG	0.545 (3)	0.63 (4)	0.58 (6)	0.37 (3)
CuriosAI	0.540 (4)	0.65 (5)	0.58 (5)	0.38 (6)
TechSSN	0.534 (5)	0.57 (14)	0.61 (2)	0.37 (4)
Sifel	0.529 (6)	0.62 (5)	0.59 (4)	0.40 (2)
BL_Research	0.505 (7)	0.59 (8)	0.54 (7)	0.33 (11)
GenAlus	0.496 (8)	0.58 (11)	0.53 (8)	0.34 (8)
ragtum	0.488 (9)	0.62 (6)	0.48 (13)	0.32 (13)
UTRAG	0.485 (10)	0.62 (7)	0.46 (11)	0.34 (10)
GigItAI	0.478 (11)	0.47 (21)	0.53 (9)	0.35 (5)
Sting	0.472 (12)	0.58 (12)	0.46 (15)	0.31 (15)
king001	0.470 (13)	0.59 (9)	0.45 (16)	0.36 (9)
PingAn	0.466 (14)	0.59 (10)	0.46 (14)	0.33 (17)
HexU	0.449 (15)	0.55 (15)	0.43 (20)	0.29 (11)
uir_cis	0.462 (16)	0.49 (18)	0.49 (10)	0.29 (18)
PFW Task 8	0.433 (17)	0.44 (22)	0.47 (12)	0.33 (12)
H-RAG	0.427 (18)	0.58 (13)	0.44 (18)	0.29 (19)
HowardU-AI4PC	0.410 (19)	0.42 (23)	0.42 (21)	0.28 (21)
Polito_Team	0.405 (20)	0.49 (17)	0.43 (19)	0.27 (22)
IITKanBDone	0.402 (21)	0.42 (24)	0.45 (17)	0.31 (16)
mt-embed	0.388 (22)	0.48 (19)	0.33 (24)	0.19 (26)
TTI-KDE	0.341 (23)	0.40 (25)	0.39 (22)	0.21 (23)
clulab-retrieval	0.331 (24)	0.51 (16)	0.30 (25)	0.28 (20)
RAID	0.285 (25)	0.47 (20)	0.26 (27)	0.20 (27)
RAGonauts	0.284 (26)	0.27 (26)	0.33 (23)	0.19 (24)
445_grads	0.229 (27)	0.25 (27)	0.26 (29)	0.07 (29)
JCT	0.174 (28)	0.21 (28)	0.19 (28)	0.10 (28)
Stratium	0.023 (29)	0.00 (29)	0.00 (29)	0.00 (29)
AKCIT-UFG	0.000 (30)	0.00 (30)	0.00 (30)	0.00 (30)

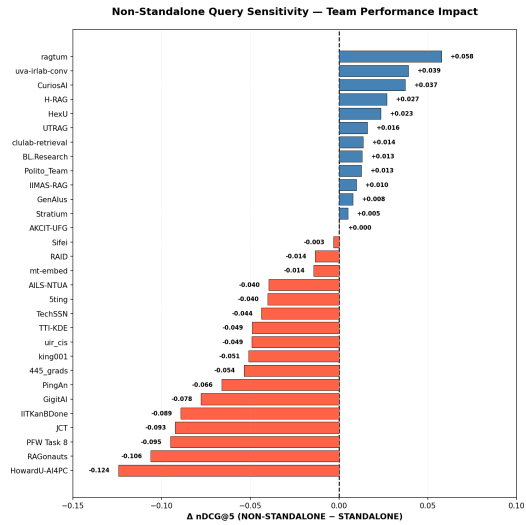


Figure 5: Δ nDCG@5 (non-standalone score minus standalone score) per team, sorted ascending. Blue bars indicate positive deltas (coreference-dependent queries help performance); red bars indicate negative deltas (performance degrades on non-standalone queries).

Figure 4: nDCG@5 scores and within-domain ranks (in parentheses) for all 30 Subtask A submissions across four retrieval domains (ClapNQ, Govt, FiQA, IBM-Cloud) plus the overall average, ordered from best to worst overall rank.

A Team Info

We provide detailed team information as described in Section 4 in Table 8.

A.1 Results

We share additional detailed results broken down by domain - Figures 4, 7, 9, answerability - Figures 6, 8, and standalone - Figure 5 for Subtasks A, B, C.

A.2 Example

We highlight an example with varying team output in Table 9.

Team (by task rank)	Answerability			
	ANSWERABLE	PARTIAL	UNANSWERABLE	UNDERSPECIFIED
RaguTeam	0.638 (1)	0.56 (2)	0.37 (10)	1.00 (1)
AILS-NTUA	0.633 (3)	0.54 (4)	0.44 (1)	1.00 (1)
YNU-HPCC	0.637 (2)	0.55 (3)	0.41 (3)	1.00 (1)
GenAlus	0.609 (6)	0.51 (10)	0.40 (6)	1.00 (1)
CuriosAI	0.587 (11)	0.49 (17)	0.36 (12)	0.99 (9)
PFW Task 8	0.608 (7)	0.52 (9)	0.33 (18)	0.99 (9)
GigItAI	0.613 (5)	0.54 (5)	0.35 (15)	0.96 (12)
Mind_Flayer	0.602 (8)	0.51 (11)	0.37 (9)	0.99 (10)
HowardU-AI4PC	0.629 (4)	0.54 (6)	0.41 (2)	1.00 (1)
RAID	0.597 (9)	0.52 (8)	0.30 (21)	0.96 (12)
ragtum	0.589 (10)	0.49 (18)	0.35 (13)	1.00 (1)
king001	0.566 (14)	0.50 (14)	0.32 (19)	0.88 (15)
IITKanBDone	0.575 (12)	0.47 (21)	0.36 (11)	1.00 (1)
Team V	0.503 (15)	0.50 (15)	0.35 (16)	0.60 (18)
UTRAG	0.545 (14)	0.56 (1)	0.40 (9)	0.57 (19)
Polito Team	0.483 (19)	0.50 (13)	0.40 (4)	0.46 (20)
Dawn	0.488 (18)	0.44 (22)	0.31 (20)	0.71 (11)
TeamRagnition	0.493 (17)	0.47 (20)	0.25 (24)	0.68 (17)
NCL	0.499 (16)	0.37 (24)	0.27 (23)	1.00 (1)
HexU	0.458 (20)	0.51 (12)	0.34 (17)	0.37 (22)
uir_cis	0.452 (21)	0.50 (16)	0.35 (14)	0.37 (22)
IMAS-RAG	0.430 (22)	0.53 (7)	0.39 (7)	0.15 (24)
uva-irlab-conv	0.368 (23)	0.48 (19)	0.39 (8)	0.04 (25)
Stratium	0.371 (24)	0.38 (23)	0.27 (22)	0.40 (14)
BL_Research	0.389 (23)	0.23 (25)	0.19 (25)	0.95 (1)

Figure 6: Mean RB_{alg} scores for Subtask B across four answerability categories (ANSWERABLE, PARTIAL, UNANSWERABLE, UNDERSPECIFIED), ordered by overall Subtask B rank.

Team (by task rank)	Overall	ClapNQ	Govt	FiQA	IBMCloud
RaguTeam	0.638 (1)	0.48 (7)	0.52 (4)	0.55 (1)	0.62 (3)
AILS-NTUA	0.633 (3)	0.49 (5)	0.51 (8)	0.53 (3)	0.62 (6)
YNU-HPCC	0.637 (2)	0.50 (4)	0.52 (3)	0.51 (5)	0.63 (2)
GenAlus	0.609 (6)	0.46 (10)	0.51 (6)	0.47 (10)	0.60 (8)
CuriosAI	0.587 (11)	0.57 (1)	0.54 (1)	0.55 (2)	0.63 (1)
PFW Task 8	0.608 (7)	0.46 (9)	0.51 (9)	0.48 (9)	0.62 (5)
GigItAI	0.613 (5)	0.47 (8)	0.52 (5)	0.49 (8)	0.60 (9)
Mind_Flayer	0.602 (8)	0.51 (3)	0.50 (10)	0.51 (4)	0.61 (7)
HowardU-AI4PC	0.629 (4)	0.49 (6)	0.51 (7)	0.51 (6)	0.62 (4)
RAID	0.597 (9)	0.52 (2)	0.52 (2)	0.50 (7)	0.60 (10)
ragtum	0.589 (10)	0.45 (12)	0.49 (11)	0.47 (11)	0.58 (13)
king001	0.566 (13)	0.43 (15)	0.46 (13)	0.43 (14)	0.58 (12)
IITKanBDone	0.575 (12)	0.45 (11)	0.47 (12)	0.44 (13)	0.59 (11)
Team V	0.503 (15)	0.41 (16)	0.43 (16)	0.40 (16)	0.49 (17)
UTRAG	0.545 (14)	0.43 (14)	0.45 (14)	0.44 (12)	0.52 (16)
Polito Team	0.483 (19)	0.36 (20)	0.40 (18)	0.37 (18)	0.49 (18)
Dawn	0.488 (18)	0.36 (21)	0.40 (19)	0.42 (15)	0.49 (19)
TeamRagnition	0.493 (17)	0.38 (19)	0.41 (17)	0.37 (17)	0.53 (14)
NCL	0.499 (16)	0.38 (18)	0.40 (20)	0.36 (19)	0.53 (15)
HexU	0.458 (20)	0.35 (22)	0.40 (21)	0.33 (23)	0.45 (20)
uir_cis	0.452 (21)	0.39 (17)	0.37 (22)	0.34 (21)	0.45 (21)
IIMAS-RAG	0.430 (22)	0.35 (23)	0.37 (23)	0.34 (22)	0.39 (23)
uva-irlab-conv	0.368 (25)	0.34 (24)	0.33 (24)	0.31 (24)	0.33 (25)
Stratium	0.371 (24)	0.34 (25)	0.26 (25)	0.28 (25)	0.37 (24)
BL_Research	0.389 (23)	0.44 (13)	0.45 (15)	0.35 (20)	0.43 (22)

Figure 7: Mean RB_{alg} scores and within-domain ranks for Subtask B (single-turn grounded generation) submissions across four domains (ClapNQ, Govt, FiQA, IBMCloud) plus overall, ordered by overall Subtask B rank.

Team (by task rank)	Overall	ANSWERABLE	PARTIAL	UNANSWERABLE	UNDERSPECIFIED
GenAlus	0.434 (3)	0.43 (7)	0.29 (13)	0.52 (6)	0.00 (13)
HowardU-AI4PC	0.443 (1)	0.45 (5)	0.29 (12)	0.49 (7)	0.05 (6)
Sting	0.387 (13)	0.43 (6)	0.34 (6)	0.28 (18)	0.03 (8)
DUTIR	0.407 (7)	0.41 (11)	0.27 (16)	0.45 (10)	0.00 (13)
GigitAI	0.438 (2)	0.39 (17)	0.21 (20)	0.69 (3)	0.00 (13)
TTI-KDE	0.417 (5)	0.40 (14)	0.21 (21)	0.56 (5)	0.00 (13)
CuriosAI	0.375 (17)	0.43 (8)	0.30 (10)	0.26 (20)	0.21 (5)
king001	0.391 (11)	0.40 (15)	0.30 (11)	0.40 (27)	0.27 (4)
PingAn	0.390 (12)	0.40 (16)	0.31 (9)	0.40 (21)	0.00 (13)
ragtum	0.380 (16)	0.41 (12)	0.31 (8)	0.32 (16)	0.00 (13)
AILS-NTUA	0.400 (9)	0.46 (4)	0.32 (7)	0.27 (19)	0.04 (7)
Georgetown CS IR Lab	0.409 (6)	0.49 (1)	0.41 (1)	0.16 (23)	0.00 (13)
IIMAS-RAG	0.399 (10)	0.47 (3)	0.28 (2)	0.19 (22)	0.00 (13)
PFW Task 8	0.420 (4)	0.39 (18)	0.20 (22)	0.62 (4)	0.01 (9)
Sifel	0.384 (15)	0.41 (13)	0.28 (14)	0.36 (14)	0.00 (13)
UTRAG	0.403 (8)	0.49 (2)	0.37 (3)	0.15 (24)	0.00 (13)
HexU	0.386 (14)	0.42 (9)	0.35 (4)	0.31 (17)	0.00 (13)
RAID	0.361 (19)	0.37 (19)	0.28 (15)	0.37 (13)	0.00 (13)
uva-irlab-conv	0.320 (22)	0.42 (10)	0.34 (5)	0.02 (27)	0.01 (9)
Polito Team	0.333 (20)	0.37 (21)	0.26 (18)	0.25 (21)	0.00 (13)
Model Citizens	0.367 (18)	0.24 (24)	0.10 (27)	0.86 (1)	0.00 (13)
IITKanBDone	0.310 (23)	0.37 (20)	0.26 (17)	0.15 (24)	0.01 (9)
NLP-CEIA-UFG	0.271 (24)	0.21 (26)	0.17 (25)	0.49 (7)	0.44 (3)
H-RAG	0.249 (26)	0.33 (22)	0.26 (19)	0.02 (27)	0.73 (2)
TechSSN	0.244 (27)	0.19 (27)	0.11 (26)	0.48 (9)	0.01 (9)
MingRAG	0.288 (23)	0.16 (28)	0.09 (28)	0.75 (2)	0.00 (13)
BL_Research	0.253 (25)	0.23 (25)	0.20 (24)	0.34 (15)	0.91 (1)
Stratium	0.217 (28)	0.25 (23)	0.20 (23)	0.13 (26)	0.00 (13)

Figure 8: Mean RB_{alg} scores for Subtask C across four answerability categories (ANSWERABLE, PARTIAL, UNANSWERABLE, UNDERSPECIFIED), ordered by overall Subtask C rank.

Team (by task rank)	collection			
	overall	cs100	csmt	POC @libcloud
GenAlus	0.434 (3)	0.32 (11)	0.37 (6)	0.31 (5)
HowardU-AI4PC	0.443 (1)	0.38 (2)	0.37 (5)	0.27 (15)
5ting	0.387 (13)	0.32 (9)	0.35 (12)	0.26 (18)
DUTIR	0.407 (7)	0.33 (8)	0.35 (13)	0.30 (9)
GigItAI	0.438 (2)	0.29 (19)	0.38 (3)	0.34 (1)
TTI-KDE	0.417 (5)	0.31 (14)	0.37 (9)	0.32 (4)
CuriosAI	0.375 (17)	0.36 (5)	0.38 (2)	0.28 (13)
king001	0.391 (11)	0.39 (1)	0.39 (1)	0.27 (16)
PingAn	0.390 (12)	0.31 (16)	0.35 (11)	0.25 (21)
ragtum	0.380 (16)	0.31 (15)	0.34 (17)	0.25 (22)
AILS-NTUA	0.460 (9)	0.34 (7)	0.35 (10)	0.32 (3)
Georgetown CS IR Lab	0.409 (6)	0.31 (12)	0.37 (7)	0.33 (2)
IIMAS-RAG	0.399 (10)	0.31 (13)	0.37 (8)	0.31 (7)
PFW Task 8	0.420 (4)	0.32 (10)	0.34 (15)	0.29 (11)
Sifei	0.384 (15)	0.30 (17)	0.35 (14)	0.30 (10)
UTRAG	0.401 (8)	0.34 (6)	0.33 (19)	0.31 (6)
HexU	0.386 (14)	0.29 (22)	0.33 (18)	0.28 (12)
RAID	0.361 (19)	0.29 (20)	0.33 (20)	0.27 (17)
uva-irlab-conv	0.320 (21)	0.28 (23)	0.27 (25)	0.28 (20)
Polito Team	0.331 (20)	0.29 (21)	0.29 (23)	0.23 (25)
Model Citizens	0.367 (18)	0.26 (24)	0.31 (22)	0.24 (23)
IITKanBDone	0.310 (22)	0.22 (26)	0.28 (24)	0.26 (19)
NLP-CEIA-UFG	0.271 (24)	0.30 (18)	0.31 (21)	0.21 (27)
H-RAG	0.249 (26)	0.38 (3)	0.34 (16)	0.27 (14)
TechSSN	0.244 (27)	0.19 (27)	0.17 (27)	0.22 (26)
MingRAG	0.288 (23)	0.23 (25)	0.21 (26)	0.24 (24)
BL_Research	0.253 (25)	0.37 (4)	0.37 (6)	0.31 (8)
Stratium	0.217 (28)	0.17 (28)	0.16 (28)	0.20 (28)

Figure 9: Mean RB_{alg} scores and within-domain ranks for Subtask C (multi-turn grounded generation) submissions across four domains plus overall, ordered by overall Subtask C rank.

Team Name	Affiliation	Academia/ Industry	Subtasks	Q	Paper
445_grads	Sabancı University	Academia	A		✓
5ting	University of Information Technology, Ho Chi Minh City, Vietnam	Academia	A	C	✓ (T-Nguyen et al., 2026)
AILS-NTUA	National Technical University of Athens	Academia	A B C		✓ (Athanasidou et al., 2026)
AKCIT - UFG	Federal University of Goiás	Academia	A		✓ (Ferreira et al., 2026)
BL.Research	Berger-Levrault	Industry	A B C		✓
caraman	Babes Bolyai University	Academia	A		✓ (Caraman and Silaghi, 2026)
clublab-retrieval	University of Arizona	Academia	A		✓ (Kim et al., 2026)
CuriosAI	SoftBank Corporation	Industry	A B C		✓ (Kumar et al., 2026)
Dawn	Yunnan University	Academia	B		✓ (Li et al., 2026b)
DUTIR	Dalian University of Technology	Academia		C	✓ (Yang et al., 2026)
evalx	IIIT Hyderabad	Academia	A	C	✗
GenAIus	GenAIus Technologies	Industry	A B C		✓ (Yeniterzi and Yeniterzi, 2026)
GigitAI	GigitAI	Industry	A B C		✓ (Krishnasamy and Wihardjo, 2026)
GUIR	Georgetown University	Academia	A	C	✓ (Abrishamchian et al., 2026)
H-RAG	Ulm University, Johannes Kepler University	Both	A	C	✓ (Elchafei et al., 2026)
HexU	Chalmers University of Technology	Academia	A B C		✓
Howard University-AI4PC	Howard University	Academia	A B C		✓ (Shrestha and Aryal, 2026)
IIMAS-RAG	Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México (UNAM)	Both	A B C		✓ (Raya-Rios et al., 2026)
IITKanBDone	IIT Kanpur, India	Both	A B C		✓ (Ray and Gupta, 2026)
JCT	JCT - Jerusalem College of Technology	Both	A		✓ (Farhan and Liebeskind, 2026)
king001	PingAn	Industry	A B C		✓
kkanbu	Korea University	Academia	A		✗
Mind_Flayer	Rajshahi University of Engineering & Technology	Academia	B		✓ (Tuli et al., 2026)
MingRAG	University of Arizona	Academia		C	✓
Model Citizens	Intapp Inc.	Industry		C	✓
mt-embed	IBM	Industry	A		✓
NCL	Newcastle University	Academia	B		✓ (Liu and Liang, 2026)
NLP-CEIA-UFG	Federal University of Goiás	Academia	A	C	✓ (Dutra et al., 2026)
PFW Task 8	Purdue University Fort Wayne	Academia	A B C		✓ (Tamsal, 2026)
PingAn	N/A	Industry	A	C	✓
Polito Team	Politecnico di Torino	Academia	A B C		✓ (Çelik et al., 2026)
RAGonauts	Sri Sivasubramaniya Nadar College of Engineering	Academia	A		✓ (Sivanaiah et al., 2026)
ragtum	TUM	Academia	A B C		✓ (Wigger et al., 2026)
RaguTeam	Novosibirsk State University	Academia	B		✓ (Derunets et al., 2026)
RAID	Barclays Investment Bank	Industry	A B C		✓
Sifei	HSE University	Academia	A	C	✓ (Meng and Ilvovsky, 2026)
SlugRAG	UC Santa Cruz	Academia	A		✓ (Revankar et al., 2026)
Stratium	University of Augsburg, Germany	Academia	A B C		✓
TDSA_UIT	University of Information Technology - VNUHCM	Academia	A		✗
Team V	Oogwai Analytics, Backbase	Industry	B		✓
TeamRagnition	Jahangirnagar University	Academia	B		✓
TechSSN	Sri Sivasubramaniya Nadar College of Engineering, India	Academia	A	C	✓ (J et al., 2026)
TTI-KDE	Toyota Technological Institute	Academia	A	C	✓
uir_cis	University of International Relations	Academia	A B		✓ (Zhang et al., 2026)
UTRAG	University of Tübingen	Academia	A B C		✓ (Zhou and Lin, 2026)
uva-irlab-conv	University of Amsterdam	Academia	A B C		✓ (Lupart et al., 2026)
VerbanexAI	Universidad tecnológica de bolívar	Academia	B		✓
YNU-HPCC	Yunnan University	Academia	A B		✓ (Li et al., 2026a)

Table 8: Detailed participating team information. Q captures if a team submitted the final questionnaire with information about their team and solution.

Conversation
<p>User: How old do I have to be to drive in CA?</p> <p>Agent: To get a provisional instruction permit or the "learner's permit", you must be at least 15 ½ years old. However, I do not have information on the minimum age requirements for a driver's license in California.</p> <p>User: Are you sure? I am pretty sure you can get a permit as soon as you are 15 years old.</p> <p>Agent: Based on my information, the minimum age to get a provisional instruction permit in California is 15 ½ years.</p> <p>User: How do I apply?</p> <p>Agent: The steps differ depending on whether you want to apply for a California Instruction Permit or a California Motorcycle Instruction Permit. If you tell me which type of permit you are interested in, I can provide more information.</p> <p>User: The regular one.</p> <p>Agent: If you are under age 18, to apply for a California Instruction Permit, you must: - Be at least 15 ½ but under 18. - Have a Certificate of Completion/Enrollment of Driver Education. - Complete the California Driver's License or ID Card Application. - Have a parent(s)/guardian(s) sign the application. - Visit a DMV office (schedule an appointment), where you will: - Bring your California Identification Card or proof of identity and residency. (Federal Noncompliant/REAL ID/AB60) - Pay the application fee. - Pass a vision exam. - Take a photo. - Take the knowledge test(s). There are multiple choice questions based on the California Driver's Handbook, and a passing score is 80%. If you are over the age of 18, the steps differ. Please let me know if you want more details about that process.</p> <p>User: The Driver's License or ID Card Application and fee never expire, correct?</p>
Relevant Passage
<p>Instruction and Learner's Permits - California DMV The letter must be signed by a school official stating that the completed courses you have taken are equivalent to a California secondary school course as described in Section 10020 of Title 5 of the California Code of Regulations and on form "To Secondary Schools Other Than California Schools" (DL 33). Obtain a DL 33 form from your local DMV office or by calling DMV at 1-800-777-0133. Send the DL 33 form to your out-of-state school for completion. An out-of-state driving instruction permit is not acceptable proof of DE/DT. What if my provisional instruction permit expires before I take the driving test? The California Driver's License or ID Card Application and fee is valid for 12 months after you apply. If your application expires before you have held your provisional instruction permit for the required six months to take your driving test, you'll need to reapply at a DMV office. You will get cumulative credit for the time you held your instruction permit that counts towards your six-month requirement. Make sure to bring both your expired and valid provisional instruction permits when you take your driving test to show proof of credit.</p>
Responses
Reference Gold
<p>Wrong. The California Driver's License or ID Card Application and fee is valid for 12 months after you apply.</p> <p>Correct Team Answer (RB_{llm}: 0.95, RB_{alg}: 0.7)</p>
<p>No, that is not correct. The California Driver's License or ID Card Application and fee is valid for 12 months after you apply. This means both the application and its payment expire exactly one year from the date they are submitted.</p> <p>Correct Team Answer (RB_{llm}: 1, RB_{alg}: 0.77)</p>
<p>No. The documents state that the California Driver's License or ID Card Application and fee is valid for **12 months after you apply**; if it expires, you'll need to reapply at a DMV office.</p> <p>Incorrect IDK Team Answer (RB_{llm}: 0, RB_{alg}: 0)</p>
<p>I do not know the answer to whether the Driver's License or ID Card Application and fee expire, as the provided context does not contain this information.</p> <p>Incorrect IDK Team Answer (RB_{llm}: 0, RB_{alg}: 0)</p>
<p>I apologize, but I'm not able to answer this question based on the provided information.</p> <p>Incorrect Irrelevant Team Answer (RB_{llm}: 0.15, RB_{alg}: 0.34)</p>
<p>In California, you must be at least 15 ½ years old to get a provisional instruction permit, also known as a learner's permit. To apply, you need to complete the California Driver's License or ID Card Application, have a parent or guardian sign it, and bring it, along with proof of identity and residency, to a DMV office. You will also need to pay the application fee, pass a vision exam, take a photo, and pass the knowledge test(s). The knowledge test consists of multiple choice questions based on the California Driver's Handbook, and a passing</p> <p>Correct Team Answer with extra information (RB_{llm}: 0.95, RB_{alg}: 0.53)</p>
<p>The California Driver's License or ID Card Application and its fee are **not** permanent. According to the DMV information, the application and fee remain valid for **12 months after you apply**. If your provisional instruction permit expires before you take the driving test, you'll need to re-apply at a DMV office. The time you held your instruction permit before it expired counts toward the required six-month period, but the application itself must be renewed...</p>

Table 9: A conversation from the govt domain that was annotated in the human eval where there was high disagreement across teams. We highlight some team answers from Subtask C with the reference answer. The relevant passage is shortened for space. In this example, many teams got it correct (8 teams), while others said I don't know (9 teams). Some also had extra information or were completely incorrect. (8 teams)