MMU-RAGent: Massive Multi-Modal User-Centric Retrieval Augmented Generation Benchmark

Luo Qi Chan[1] CLUOQI@DSO.ORG.SG
Tevin Wang[1]
Shuting Wang[1]
Zhihan Zhang[1]
Alfredo Gomez[1]
Prahaladh Chandrahasan[1]
Andy Tang[1]
Lan Yan[1]
Zimeng (Chris) Qiu[2]
Morteza Ziyadi[2]
Tongshuang Wu[1]
Mona Diab[1]
Akari Asai[1]
Chenyan Xiong[1]

Izat Temiraliev[3] and **Serene Cheng**[3] and **Diji Yang**[3] and **Yi Zhang**[3]
Kun Ran[4] and **Marwah Alaofi**[5] and **Danula Hettiachchi**[4] and **Chenglong Ma**[4] and **Khoi Nguyen Dinh Anh**[4] and **Khoi Vo Nguyen**[4] and **Sachin Pathiyan Cherumanal**[4] and **Lida Rashidi**[4] and **Falk Scholer**[4] and **Damiano Spina**[4] and **Shuoqi Sun**[4] and **Oleg Zendel**[4]
Ikuya Yamada[6,7] and **Wataru Ikeda**[7] and **Ko Yoshida**[7] and **Mengyu Ye**[7] and **Hinata Sugimoto**[7] and **Masatoshi Suzuki**[6,7] and **Hisanori Ozaki**[8] and **Jun Suzuki**[7,9,10]
Cattalyya Nuengsigkapian[11]
Dmitrii Magas[12]
Quentin Fever[13] and **Naziha Aslam**[14]

[1] Work done at Carnegie Mellon University, Pittsburgh, USA. Organizing Team.
[2] Amazon AGI. Organizing Team.
[3] University of California, Santa Cruz, USA. Participated as `Team DeepVideoResearcher`.
[4] RMIT University, Melbourne, Australia. Participated as `Team RMIT-ADMS IR`.
[5] Taibah University, Medina, Saudi Arabia. Participated as part of `Team RMIT-ADMS IR`.
[6] Studio Ousia. Participated as part of `Team efficient-deep-research`.
[7] Tohoku University. Participated as part of `Team efficient-deep-research`.
[8] DENTSU SOKEN INC.Participated as part of `Team efficient-deep-research`.
[9] RIKEN. Participated as part of `Team efficient-deep-research`.
[10] NII LLMC. Participated as part of `Team efficient-deep-research`.
[11] Google. Participated as `Team Cattalyya`.
[12] Independent Researcher. Participated as `Team TTT-DR`.
[13] SESAMm. Participated as part of `Team NightFeats`.

© 2026 Chan *et al.*

[14] GMV (Grupo Mecánica del Vuelo). Participated as part of `Team NightFeats`.
Superscripts denote affiliations only and do not indicate author contribution or ordering.

Abstract

We present the results of the MMU-RAGent competition. The competition tasked participants with building systems capable of answering real-world queries drawn from MS MARCO Web Search and Chatbot Arena conversations, as well as queries collected live through the RAG-Arena evaluation platform. To support large-scale retrieval, participants were given API access to the English subsets of ClueWeb22-B and ClueWeb22-A (87M and 800M documents) and AWS-hosted infrastructure for system deployment and evaluation. Submissions were evaluated using a combination of human Likert-scale ratings, live preference judgments collected via RAG-Arena, LLM-as-a-Judge evaluation, and automatic metrics. The competition allowed both open- and closed-source approaches, including systems leveraging proprietary search APIs or models, with separate leaderboard categories to ensure transparent comparison. Participating systems explored a wide range of RAG architectures, including hierarchical retrieval pipelines, structured document parsing, multi-stage reranking, and iterative generation with citation verification. The competition provides a new benchmark for studying retrieval-augmented generation under realistic deployment conditions and highlights emerging system design patterns for scalable and user-aligned RAG.

Keywords: Retrieval-Augmented Generation, Large-Language Models, Information Retrieval, Multimodality, User-centric Evaluations

1. Introduction

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Guu et al., 2020; Ram et al., 2023) has become a widely adopted approach for improving the factuality and grounding of large language models (LLMs) by incorporating external knowledge during generation (Mallen et al., 2023; Kasai et al., 2023). Yet, most existing evaluations rely on static benchmarks (Kwiatkowski et al., 2019; Wei et al., 2024, 2025), which often fail to reflect the challenges faced by real-world RAG systems, such as handling diverse user queries (Cao et al., 2026) and retrieving information from large, noisy web corpora (Wan et al., 2024).

To study RAG systems under more realistic conditions, this competition evaluates end-to-end RAG systems using real-user queries and web-scale retrieval. Participants built systems capable of answering queries drawn from MS MARCO Web Search (Chen et al., 2024), Chatbot Arena (Zheng et al., 2023) conversations, and live queries collected through the RAG-Arena platform. Systems retrieved evidence from the ClueWeb22 (Overwijk et al., 2022) corpus and generated responses in both text and video formats. Submissions were evaluated using a combination of human judgments, live preference comparisons, LLM-based evaluation, and automatic metrics.

The competition highlights emerging design patterns for large-scale RAG systems, including multi-stage retrieval pipelines, LLM-based reranking, and iterative answer generation with verification. The released datasets, evaluation framework, and competition results provide a new benchmark for studying retrieval-augmented generation in realistic deployment settings.

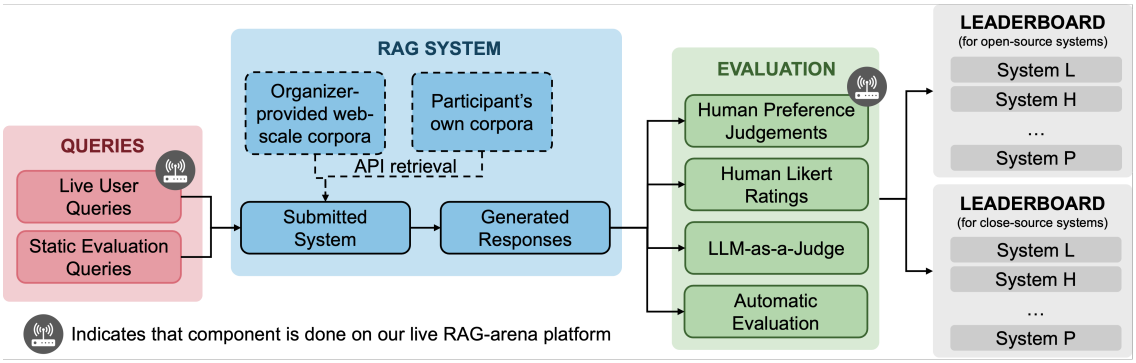


Figure 1: Competition Overview.

2. Competition Overview and Data Used.

Compared to previous RAG-related competitions, this benchmark differs in four main aspects: (1) the use of open-domain real-user queries and user-centric evaluation; (2) retrieval over web-scale corpora; (3) the introduction of a retrieval-augmented text-to-video generation task; and (4) live evaluation and query collection through the RAG-Arena platform. Figure 1 shows the competition overview.

- **Real-user queries and evaluation.** Queries are drawn from MS MARCO Web (Chen et al., 2024), Chatbot Arena Conversations (Chiang et al., 2024), and live queries collected through the RAG-Arena platform (See Appendix A for more details on data). Systems are evaluated using human Likert-scale ratings, pairwise human preference judgments, and automatic metrics. In contrast to domain-specific competitions such as FinanceRAG (Choi et al., 2024) or service-oriented evaluations such as FutureDial'24 (Cai et al., 2024), this competition focuses on open-domain information-seeking queries.
- **Web-scale retrieval.** Participants were provided high-throughput API access (100 to 400 queries per second) to the English subsets of ClueWeb22-B and ClueWeb22-A (Overwijk et al., 2022), containing approximately 87M and 800M documents, respectively. In contrast to prior benchmarks such as EfficientQA (Min et al., 2021), which uses the 5M-document KILT-Wikipedia corpus (Petroni et al., 2021), this competition enables experimentation with large-scale web retrieval. While CRAG Yang et al. (2024) provides simulated web APIs, only the top-50 documents are returned per query, limiting retrieval exploration over the full corpus.
- **Multimodal generation.** In addition to traditional text-based RAG, the competition introduces a retrieval-augmented text-to-video generation task. Existing video generation benchmarks (He et al., 2023; Wang et al., 2024) primarily evaluate visual generation from text prompts, while common automatic metrics such as VBENCH (Huang et al., 2023) and EvalCrafter Liu et al. (2024) focus on visual quality rather than factual grounding.
- **Live user evaluation.** The competition platform supports Chatbot Arena-style evaluation (Chiang et al., 2024), enabling pairwise comparisons and the collection of user queries and feedback in real time. Such live evaluation infrastructure is uncommon in academic competitions and allows systems to be assessed in an interactive setting.

Table 1: Illustration of our static evaluation methods and their corresponding metrics.

Track	Evaluation Method	Evaluation Metric
Text-to-text	Automatic	Rouge-L (Lin, 2004), BERTScore (Zhang et al., 2020)
	LLM-as-a-Judge	Semantic Similarity, Coverage, Factuality
	Human Likert Ratings	
Text-to-video	Automatic	Subject Consistency, Background Consistency, Motion Smoothness, Dynamic Degree,Aesthetic Quality, Imaging Quality (Huang et al., 2023)
	LLM-as-a-Judge	Relevance, Precision, Recall, Usefulness
	Human Likert Ratings	

3. Evaluation Protocol

This section describes our evaluation protocol and the corresponding metrics used to evaluate each track. Table 1 shows an outline of our evaluation protocol and metrics.

We use both *static* and *live* evaluations to provide a comprehensive assessment. *Static* evaluation measures performance on predefined queries, while *live* evaluation captures real-time user feedback via our RAG-Arena platform.

3.1. Static Evaluation

Static evaluation consists of three components. (1) **Automatic evaluation** uses standard metrics such as ROUGE-L (Lin, 2004) to ensure reproducibility. (2) **LLM-as-a-Judge** (Gu et al., 2025) provides high-level assessments on nuanced dimensions, using LLMs fine-tuned on human annotations as scalable proxies for human judgment. (3) **Human Likert ratings**, collected via Prolific, offer direct human evaluation and are also used to train and validate the LLM-based judgments.

Track A: Text-to-text. For automatic metrics, we select *Rouge-L* (Lin, 2004) and *BERTScore* (Zhang et al., 2020). To enhance semantic and multi-aspect evaluation, we introduce three metrics assessed by both LLM-as-a-judge and human Likert ranking methods. First, *semantic similarity* measures the alignment between generated and reference answers, accounting for paraphrasing and surface-level variation that term- or embedding-level matching tends to miss. Second, *fact coverage* (Asai et al., 2024; Wang et al., 2025) assesses the completeness of a response by checking how well it addresses the factual key points in the reference.

Track B: Text-to-video. We evaluate video responses for (1) video quality and (2) video utility. Video quality is measured using standard automatic metrics from VBench (Huang et al., 2023) as listed in Table 1.

While VBench metrics capture visual fidelity and prompt alignment, they do not evaluate whether a generated video fulfills the user's informational need. To address this gap, we conducted pilot annotations and two rounds of focused group discussions, guided by inter-annotator agreement studies, to refine a rubric that annotators found both intuitive and reliable. This process led to a two-part evaluation framework combining visual quality and task utility.

For utility, we introduce four novel metrics, assessed through LLM-as-a-judge and human Likert ratings. First, *Relevance* measures how well the video content aligns with the query. Second, *Precision* captures the degree to which the video remains focused and avoids irrelevant or incorrect material. Third, *Recall* evaluates whether the video includes all essential visual actions and elements needed to adequately address the query. Last, *Usefulness* reflects the extent to which the video effectively assists users in accomplishing the task posed by the query. The full rubric is provided in Appendix D.

To validate these dimensions as usable evaluation criteria, we conducted a final round of multi-rater annotations on the utility-focused metrics. Inter-annotator agreement, measured with Krippendorff's α, averaged 0.21 across the four metrics. This level of consistency is reasonable given the inherent subjectivity of fine-grained video utility assessment, particularly with six independent annotators. Importantly, we found that aggregate ratings were stable across annotators.

3.2. Ranking Methodology

Before computing final rankings, we first examined the relationship between LLM-as-a-judge scores and human Likert evaluations for the semantic dimensions shared across both methods. We found that LLM-based assessments were highly correlated with human ratings for both *semantic similarity* ($r = 0.926$) and *fact coverage* ($r = 0.930$). This indicates that LLM judgments largely mirror human preferences for these dimensions. To avoid double-counting the same underlying signal in the final score, we therefore exclude LLM-as-a-judge scores for these two metrics from the aggregation and rely solely on human evaluations.

To derive the final leaderboard, we aggregate automatic metrics and human Likert evaluations into a single robustness-aware score. For each system, we compute an α-weighted aggregate:

$$\text{Score}(\alpha) = \alpha \cdot \sum_{\text{auto}} \text{AutoMetric} \; + \; (1-\alpha) \cdot \sum_{\text{human}} \text{HumanLikert}.$$

All automatic metrics and human ratings are first normalized on a per-metric basis. The weighting parameter $\alpha \in [0, 1]$ determines the relative contribution of automatic versus human evaluations. To mitigate the subjectivity of this choice, we conduct a stability analysis by sweeping α from 0 to 1 and computing a ranking at each step. The final main leaderboard is defined as the most stable ordering across all α values to ensuring robustness.

3.3. Live Evaluation

Beyond static evaluation, we use (4) **Arena-style user-preference evaluation** (Chiang et al., 2024) for *live* evaluations. We host a live RAG-arena platform where users submit their queries and two randomized models generate responses in real-time. Then, users label their preferred generations in real-time. Such pairwise preferences reveal subtle qualitative distinctions that static metrics might overlook.

This multipronged approach ensures that systems are assessed not only for correctness but also for how well they align with real user expectations in practical scenarios.

4. Participation and Results

This section presents the outcomes of the static and live evaluations described in Section 3.

4.1. Participation Overview and Statistics

The 2025 MMU-RAGent Competition attracted active participation across both tracks. For the Text-to-Text track, we received a total of six full system submissions, supplemented by two additional submissions evaluated on the validation set only. The Text-to-Video track received one complete system submission. Details of each system can be found in Appendix H, and we discuss lessons from participants in Section 5.

To support development and benchmarking, we released multiple query sets throughout the competition. Participants were initially provided with a 300-query development set. Subsequently, a 299-query validation set was released closer to the submission deadline, enabling teams to perform held-out evaluation. Final rankings were computed using a 400-query unseen test set, which remained hidden throughout the competition.

Human evaluation constituted a critical component of our assessment pipeline. Across both tracks, we collected 2,315 individual annotations from 1,197 annotators. For the Text-to-Text track, we gathered 2,265 total ratings. For the Text-to-Video track, we collected 50 utility-focused annotations across relevance, precision, recall, and usefulness.

4.2. Results of Static Evaluation

We benchmark two proprietary large-model baselines, Claude-SonnetV2 and Nova-Pro, under the same static evaluation protocol. These baselines provide an upper-bound reference for semantic similarity and lexical alignment in a non-retrieval setting. While they generally achieve strong automatic metric scores, they underperform in LLM-as-a-judge evaluations compared with the top RAG systems.

Automatic Metrics. Across both proprietary baselines and participant systems, BERTScore values are tightly clustered (0.82–0.85), while ROUGE-L exhibits larger variation, reflecting differences in lexical expressiveness. Claude-SonnetV2 and Nova-Pro achieve strong BERTScores but relatively lower ROUGE-L compared to the top RAG systems. Among submitted teams, `slugresearcher` and `RMIT-ADMS_IR` obtain the strongest automatic scores, suggesting strong lexical and semantic alignment with reference answers.

LLM-as-a-Judge. LLM-based semantic similarity and fact coverage reveal a wider separation between systems than automatic metrics. The strongest teams (`RMIT-ADMS_IR`, `efficient-deep-research`) achieve both high semantic alignment and strong factual grounding. The proprietary baselines sit mid-spectrum: Claude-SonnetV2 approaches the top systems in semantic similarity but underperforms in factual coverage, while Nova-Pro lags on both metrics. This contrast highlights the benefit of retrieval-augmented reasoning for producing well-grounded answers. Lower-ranked systems show consistent drops across both LLM-judged dimensions.

Human Likert Evaluations. Human evaluations serve as the primary semantic and factual signal in the final ranking. Ratings for the top four systems show clear separation:

Table 2: Baselines from proprietary systems and top system performances across metrics.

Team Name	Automatic		LLM-as-a-Judge		Human Likert	
	Rouge-L	BERTScore	Sem. Sim.	Fact Cov.	Sem. Sim.	Fact Cov.
claude-sonnetV2 (Baseline)	0.181	0.850	0.809	0.572	-	-
nova-pro (Baseline)	0.200	0.851	0.670	0.432	-	-
`efficient-deep-research`	0.111	0.837	0.820	0.571	3.735	4.035
`RMIT-ADMS_IR`	0.133	0.825	0.809	0.607	3.825	4.030
`slugresearcher`	0.135	0.831	0.612	0.375	3.305	3.710
`TTT-DR`	0.042	0.832	0.705	0.536	3.670	4.065

Table 3: RAG-Arena live evaluation results for the top four Text-to-Text systems.

Team Name	RAG-Arena Score
`RMIT-ADMS_IR`	1204.77
`TTT-DR`	1134.10
`efficient-deep-research`	1025.86
`slugresearcher`	975.12

Table 4: Final static ranking after aggregation (Section 3.1).

Ranking	Team Name
1	`efficient-deep-research`
2	`RMIT-ADMS_IR`
3	`TTT-DR`
4	`slugresearcher`

`RMIT-ADMS_IR` and `efficient-deep-research` receive the highest semantic similarity and coverage scores, followed closely by `TTT-DR`, with `slugresearcher` showing comparatively lower but still competitive performance.

Final Static Ranking. Final rankings were computed via the robustness-aware aggregation procedure described in Section 3.1, combining normalized automatic metrics and human evaluations using an α-weighted score. Table 4 shows the top four systems selected for live evaluation. These rankings were stable across the full sweep of $\alpha \in [0, 1]$.

4.3. Live Evaluation

We conducted live user evaluations for the top four Text-to-Text systems using our RAG-Arena platform. In this setting, models are compared pairwise on unseen user queries, and users select their preferred responses. Table 3 reports the aggregated arena scores.

While `efficient-deep-research` was the most stable system under the static aggregation procedure, it was outperformed by `RMIT-ADMS_IR` in the arena when judged interactively. These differences underscore the value of including a live preference-based evaluation: while static metrics capture semantic fidelity and factual grounding, arena judgments surface qualitative distinctions that matter in actual use, such as clarity and perceived utility.

4.4. Text-to-Video

The Text-to-Video track received one full-system submission (`DeepVideoResearcher`). We compare the submitted system against a baseline video generator built using Nova-Reel (AGI et al., 2025), prompted directly with the user query. We aggregate rankings across metrics using Borda scores to enable direct comparison between models. In this framework, systems

Table 5: Automatic VBench evaluation for the Text-to-Video track.

Team	Subj. Cons.	Back. Cons.	Motion	Dynamic	Aesthetic	Imaging
Nova-Reel	0.97	0.98	0.99	0.23	0.59	0.70
`dvr`	0.79	0.90	0.99	0.90	0.52	0.62

are ranked within each metric and assigned points according to their position; the final score reflects their cumulative standing across all evaluated dimensions.

Automatic Evaluation. Table 5 reports the automatic VBench metrics for both systems. The baseline system achieves higher scores on subject consistency, background consistency, motion smoothness, aesthetic quality, and imaging quality. In contrast, `DeepVideoResearcher` (dvr) shows substantially stronger performance only in dynamic degree, indicating higher motion diversity but weaker visual stability and fidelity overall. These results suggest that the submitted system lags behind the baseline on conventional visual-quality axes.

Human Evaluation. Details of the human evaluation dimensions and annotation rubric are provided in Section 3.1 (Text-to-Video Evaluation). Despite weaker automatic scores, `DeepVideoResearcher` (3.20) outperforms the baseline (2.94) in human likert ratings across seven categories. Annotators judged its videos as more relevant, more precise, and more useful for answering the user's query. Improvements are also observed in step quality and overall utility, even though the baseline slightly outperforms on visual realism (see Appendix E).

Although the baseline system performs better on nearly all automatic VBench metrics, `DeepVideoResearcher` produces videos that humans judge to be more aligned with the informational needs of the query. This divergence highlights a central challenge for text-to-video evaluation: current automatic metrics primarily assess visual fidelity and aesthetic quality, but they do not reliably capture task relevance, procedural correctness, or utility. These results reinforce the importance of incorporating human-centered evaluation when assessing multimodal RAG systems.

5. Participant Methodologies and Lessons

We invited teams to share their systems and summarized the methodological patterns observed across systems. Detailed team write-ups are in Appendix H. Taken together, the submissions across both tracks suggest an emphasis on retrieval quality, evidence processing, bounded iteration, and deployment-aware design.

5.1. Text-to-Text Systems

We summarize and present systems from teams `Cattalyya`, `efficient-deep-research`, `RMIT-ADMS_IR`, `NightFeats`, and `TTT-DR` . While their architectures vary, several recurring design patterns emerge from the system descriptions and reported lessons.

Precision-Oriented Retrieval and Context Curation. Multiple teams emphasize improving the precision and structure of retrieved context prior to generation. `Cattalyya`

describes a multi-stage filtering pipeline consisting of query reformulation, URL filtering, hierarchical content parsing, LLM-based reranking, and post-hoc citation verification. The team reports that prompt engineering and content filtering were primary contributors to performance. `NightFeats` similarly decomposes its pipeline into retrieval, curation, and composition stages, incorporating hybrid retrieval, reranking, atomic fact extraction, and bounded contradiction reconciliation. These approaches treat context selection and evidence organization as central determinants of downstream answer quality.

Iterative Retrieval with Explicit Termination Criteria. Several systems implement multi-round retrieval and refinement rather than a single retrieve–then–generate pass. `RMIT-ADMS_IR` routes queries to either a single-pass RAG pipeline or an iterative agent-style pipeline that accumulates evidence across rounds. Retrieval terminates when coverage is deemed sufficient, a token budget is exceeded, or a maximum iteration count is reached. `TTT-DR` frames report generation as iterative refinement of an evolving draft document, where retrieval queries are conditioned on the current draft state. The team reports reducing iteration depth to meet evaluation time constraints. These designs reflect a shared emphasis on controlling iteration depth to balance answer quality against latency.

Bounded Agentic Components and Critique Mechanisms. Some text-to-text systems incorporate reviewer or critic components to assess evidence sufficiency or resolve contradictions. `NightFeats` employs bounded reconciliation cycles to address conflicting facts, while `TTT-DR` uses LLM-based self-critique for intermediate planning and synthesis steps. Teams report practical challenges in tuning critique strictness and managing additional latency introduced by agentic components, motivating explicit iteration limits and simplified control flows.

Training-Time Optimization for Long-Form Generation. `efficient-deep-research` focuses on training-time preference optimization using synthetic data and Direct Preference Optimization (DPO) with LLM-as-a-judge signals. The team reports strong performance on long-form research-style outputs without prior supervised fine-tuning, but notes a tendency toward longer responses, which may be less suitable for short factoid queries. This approach contrasts with inference-centric orchestration strategies adopted by other teams.

Engineering and Deployment Constraints. Several teams explicitly describe deployment-driven design decisions. `RMIT-ADMS_IR` and `TTT-DR` report tuning vLLM configurations, context lengths, and reranker integration to operate within single-GPU and containerized environments. `NightFeats` highlights latency and concurrency challenges arising from multi-agent orchestration and external API calls. These accounts indicate that system engineering considerations, such as memory allocation, throughput, and packaging constraints, substantially shaped practical text-to-text RAG designs.

5.2. Text-to-Video Systems

One team (`DeepVideoResearcher`) submitted a text-to-video system. Compared to text-to-text pipelines, the system architecture reflects distinct cost and control-flow constraints associated with video generation.

Deferred Generation with Upstream Validation. `DeepVideoResearcher` treats video synthesis as the final step of a research-oriented pipeline. User prompts are decomposed into structured research questions, followed by iterative retrieval and evidence verification. An external Critic Agent evaluates whether retrieved evidence is sufficient, consistent, and aligned with user intent before invoking the diffusion-based video generator (Yang et al., 2025). The team reports that this design was motivated by need to decouple iterative grounding from heavy-compute synthesis, considering high runtime cost of video generation.

Critic Calibration and Control Flow Challenges. The system relies on an external critique loop to prevent hallucinated or insufficient grounding. However, the team reports challenges in calibrating critical strictness: overly strict critique increased runtime by repeatedly requesting additional evidence, while lenient critique allowed subtle gaps in visual grounding to propagate to the final output. These observations highlight the sensitivity of multimodal RAG systems to control-flow design when generation is computationally expensive.

Optimization at the Text–Video Interface. Rather than modifying the video generator itself, `DeepVideoResearcher` concentrates optimization on the alignment between the research output and the generator-ready prompt. The team notes that while this approach is effective under a fixed generator assumption, further improvements may require joint optimization or fine-tuning of the video model.

6. Evaluation and Annotation Feedback

In this section, we analyze annotator feedback from 219 pairs of evaluation instances from RAGArena to better understand the strengths and limitations of current DR-style systems and their evaluation. Beyond aggregate performance metrics, this analysis provides insight into how annotators perceive response quality in practice, and highlights key gaps between existing evaluation paradigms and human preferences.

Annotator feedback reveals several systematic issues. First, annotators consistently noted a mismatch between response length and utility: longer, more verbose answers were not necessarily preferred, with concise and directly relevant responses often judged as higher quality. This suggests that existing systems may over-optimize for completeness at the expense of information density.

Second, structure and clarity emerged as key drivers of preference. Well-organized, easy-to-follow responses were favored even when content overlap was high. This indicates that evaluation setups should consider presentation quality.

Third, annotators distinguished between shallow retrieval and deeper reasoning. Responses that merely listed or restated information were rated lower than those that synthesized evidence or provided clear justification, highlighting the need for evaluation frameworks that better capture reasoning depth rather than factual coverage alone.

Finally, annotators frequently reported difficulty in making binary pairwise judgments, especially when responses exhibited different strengths or when both were similarly strong or weak. This introduces noise into preference-based evaluation and suggests the need for more multi-dimensional or criteria-based evaluation schemes that explicitly assess relevance, clarity, and reasoning.

Overall, these findings indicate that current evaluation paradigms do not fully align with human preferences, and point toward the need for more nuanced evaluation frameworks that capture relevance, structure, and depth of reasoning in addition to correctness.

7. Conclusion

This competition provides a holistic evaluation of RAG systems across both text and video modalities. Through a combination of static evaluation and the RAG-Arena platform, we assessed systems not only for semantic correctness and factual grounding but also for their behavior in realistic interactive settings.

Participation across tracks demonstrated strong community interest, and the diversity of submitted systems revealed clear patterns in current RAG capabilities. In the text-to-text track, static and live evaluations were largely aligned, while small differences highlighted the importance of preference-based assessments. The text-to-video track underscored the gap between visual quality and task usefulness, pointing to promising directions for future multimodal RAG research.

We hope that datasets, and evaluation methodologies introduced in this competition will support future work on building more robust, transparent, and user-aligned RAG systems. We also anticipate that the lessons learned will inform the design of next-generation benchmarks.

Acknowledgments

We express our sincere gratitude to all participants and the NeurIPS organizing committee for their support in making this competition possible. We also gratefully acknowledge Amazon AGI for their generous contribution of AWS storage and access to model APIs, without which this work would not have been possible.

References

Amazon AGI, Aaron Langford, Aayush Shah, Abhanshu Gupta, Abhimanyu Bhatter, Abhinav Goyal, Abhinav Mathur, Abhinav Mohanty, Abhishek Kumar, Abhishek Sethi, Abi Komma, Abner Pena, Achin Jain, Adam Kunysz, Adam Opyrchal, Adarsh Singh, Aditya Rawal, Adok Achar Budihal Prasad, Adrià de Gispert, Agnika Kumar, Aishwarya Aryamane, Ajay Nair, Akilan M, Akshaya Iyengar, Akshaya Vishnu Kudlu Shanbhogue, Alan He, Alessandra Cervone, Alex Loeb, Alex Zhang, Alexander Fu, Alexander Lisnichenko, Alexander Zhipa, Alexandros Potamianos, Ali Kebarighotbi, Aliakbar Daronkolaei, Alok Parmesh, Amanjot Kaur Samra, Ameen Khan, Amer Rez, Amir Saffari, Amit Agarwalla, Amit Jhindal, Amith Mamidala, Ammar Asmro, Amulya Ballakur, Anand Mishra, Anand Sridharan, Anastasiia Dubinina, Andre Lenz, Andreas Doerr, Andrew Keating, Andrew Leaver, Andrew Smith, Andrew Wirth, Andy Davey, Andy Rosenbaum, Andy Sohn, Angela Chan, Aniket Chakrabarti, Anil Ramakrishna, Anirban Roy, Anita Iyer, Anjali Narayan-Chen, Ankith Yennu, Anna Dabrowska, Anna Gawlowska, Anna Rumshisky, Anna Turek, Anoop Deoras, Anton Bezruchkin, Anup Prasad, Anupam Dewan, Anwith Kiran, Apoorv Gupta, Aram Galstyan, Aravind Manoharan, Arijit

Biswas, Arindam Mandal, Arpit Gupta, Arsamkhan Pathan, Arun Nagarajan, Arushan Rajasekaram, Arvind Sundararajan, Ashwin Ganesan, Ashwin Swaminathan, Athanasios Mouchtaris, Audrey Champeau, Avik Ray, Ayush Jaiswal, Ayush Sharma, Bailey Keefer, Balamurugan Muthiah, Beatriz Leon-Millan, Ben Koopman, Ben Li, Benjamin Biggs, Benjamin Ott, Bhanu Vinzamuri, Bharath Venkatesh, Bhavana Ganesh, Bhoomit Vasani, Bill Byrne, Bill Hsu, Bincheng Wang, Blake King, Blazej Gorny, Bo Feng, Bo Zheng, Bodhisattwa Paul, Bofan Sun, Bofeng Luo, Bowen Chen, Bowen Xie, Boya Yu, Brendan Jugan, Brett Panosh, Brian Collins, Brian Thompson, Can Karakus, Can Liu, Carl Lambrecht, Carly Lin, Carolyn Wang, Carrie Yuan, Casey Loyda, Cezary Walczak, Chalapathi Choppa, Chandana Satya Prakash, Chankrisna Richy Meas, Charith Peris, Charles Recaido, Charlie Xu, Charul Sharma, Chase Kernan, Chayut Thanapirom, Chengwei Su, Chenhao Xu, Chenhao Yin, Chentao Ye, Chenyang Tao, Chethan Parameshwara, Ching-Yun Chang, Chong Li, Chris Hench, Chris Tran, Christophe Dupuy, Christopher Davis, Christopher DiPersio, Christos Christodoulopoulos, Christy Li, Chun Chen, Claudio Delli Bovi, Clement Chung, Cole Hawkins, Connor Harris, Corey Ropell, Cynthia He, DK Joo, Dae Yon Hwang, Dan Rosen, Daniel Elkind, Daniel Pressel, Daniel Zhang, Danielle Kimball, Daniil Sorokin, Dave Goodell, Davide Modolo, Dawei Zhu, Deepikaa Suresh, Deepti Ragha, Denis Filimonov, Denis Foo Kune, Denis Romasanta Rodriguez, Devamanyu Hazarika, Dhananjay Ram, Dhawal Parkar, Dhawal Patel, Dhwanil Desai, Dinesh Singh Rajput, Disha Sule, Diwakar Singh, Dmitriy Genzel, Dolly Goldenberg, Dongyi He, Dumitru Hanciu, Dushan Tharmal, Dzmitry Siankovich, Edi Cikovic, Edwin Abraham, Ekraam Sabir, Elliott Olson, Emmett Steven, Emre Barut, Eric Jackson, Ethan Wu, Evelyn Chen, Ezhilan Mahalingam, Fabian Triefenbach, Fan Yang, Fangyu Liu, Fanzi Wu, Faraz Tavakoli, Farhad Khozeimeh, Feiyang Niu, Felix Hieber, Feng Li, Firat Elbey, Florian Krebs, Florian Saupe, Florian Sprünken, Frank Fan, Furqan Khan, Gabriela De Vincenzo, Gagandeep Kang, George Ding, George He, George Yeung, Ghada Qaddoumi, Giannis Karamanolakis, Goeric Huybrechts, Gokul Maddali, Gonzalo Iglesias, Gordon McShane, Gozde Sahin, Guangtai Huang, Gukyeong Kwon, Gunnar A. Sigurdsson, Gurpreet Chadha, Gururaj Kosuru, Hagen Fuerstenau, Hah Hah, Haja Maideen, Hajime Hosokawa, Han Liu, Han-Kai Hsu, Hann Wang, Hao Li, Hao Yang, Haofeng Zhu, Haozheng Fan, Harman Singh, Harshavardhan Kaluvala, Hashim Saeed, He Xie, Helian Feng, Hendrix Luo, Hengzhi Pei, Henrik Nielsen, Hesam Ilati, Himanshu Patel, Hongshan Li, Hongzhou Lin, Hussain Raza, Ian Cullinan, Imre Kiss, Inbarasan Thangamani, Indrayani Fadnavis, Ionut Teodor Sorodoc, Irem Ertuerk, Iryna Yemialyanava, Ishan Soni, Ismail Jelal, Ivan Tse, Jack FitzGerald, Jack Zhao, Jackson Rothgeb, Jacky Lee, Jake Jung, Jakub Debski, Jakub Tomczak, James Jeun, James Sanders, Jason Crowley, Jay Lee, Jayakrishna Anvesh Paidy, Jayant Tiwari, Jean Farmer, Jeff Solinsky, Jenna Lau, Jeremy Savareese, Jerzy Zagorski, Ji Dai, Jiacheng, Gu, Jiahui Li, Jian, Zheng, Jianhua Lu, Jianhua Wang, Jiawei Dai, Jiawei Mo, Jiaxi Xu, Jie Liang, Jie Yang, Jim Logan, Jimit Majmudar, Jing Liu, Jinghong Miao, Jingru Yi, Jingyang Jin, Jiun-Yu Kao, Jixuan Wang, Jiyang Wang, Joe Pemberton, Joel Carlson, Joey Blundell, John Chin-Jew, John He, Jonathan Ho, Jonathan Hueser, Jonathan Lunt, Jooyoung Lee, Joshua Tan, Joyjit Chatterjee, Judith Gaspers, Jue Wang, Jun Fang, Jun Tang, Jun Wan, Jun Wu, Junlei Wang, Junyi Shi, Justin Chiu, Justin Satriano, Justin Yee, Jwala Dhamala, Jyoti Bansal, Kai Zhen, Kai-Wei Chang, Kaixiang Lin, Kalyan Raman, Kanthashree Mysore

Sathyendra, Karabo Moroe, Karan Bhandarkar, Karan Kothari, Karolina Owczarzak, Karthick Gopalswamy, Karthick Ravi, Karthik Ramakrishnan, Karthika Arumugam, Kartik Mehta, Katarzyna Konczalska, Kavya Ravikumar, Ke Tran, Kechen Qin, Kelin Li, Kelvin Li, Ketan Kulkarni, Kevin Angelo Rodrigues, Keyur Patel, Khadige Abboud, Kiana Hajebi, Klaus Reiter, Kris Schultz, Krishna Anisetty, Krishna Kotnana, Kristen Li, Kruthi Channamallikarjuna, Krzysztof Jakubczyk, Kuba Pierewoj, Kunal Pal, Kunwar Srivastav, Kyle Bannerman, Lahari Poddar, Lakshmi Prasad, Larry Tseng, Laxmikant Naik, Leena Chennuru Vankadara, Lenon Minorics, Leo Liu, Leonard Lausen, Leonardo F. R. Ribeiro, Li Zhang, Lili Gehorsam, Ling Qi, Lisa Bauer, Lori Knapp, Lu Zeng, Lucas Tong, Lulu Wong, Luoxin Chen, Maciej Rudnicki, Mahdi Namazifar, Mahesh Jaliminche, Maira Ladeira Tanke, Manasi Gupta, Mandeep Ahlawat, Mani Khanuja, Mani Sundaram, Marcin Leyk, Mariusz Momotko, Markus Boese, Markus Dreyer, Markus Mueller, Mason Fu, Mateusz Górski, Mateusz Mastalerczyk, Matias Mora, Matt Johnson, Matt Scott, Matthew Wen, Max Barysau, Maya Boumerdassi, Maya Krishnan, Mayank Gupta, Mayank Hirani, Mayank Kulkarni, Meganathan Narayanasamy, Melanie Bradford, Melanie Gens, Melissa Burke, Meng Jin, Miao Chen, Michael Denkowski, Michael Heymel, Michael Krestyaninov, Michal Obirek, Michalina Wichorowska, Michał Miotk, Milosz Watroba, Mingyi Hong, Mingzhi Yu, Miranda Liu, Mohamed Gouda, Mohammad El-Shabani, Mohammad Ghavamzadeh, Mohit Bansal, Morteza Ziyadi, Nan Xia, Nathan Susanj, Nav Bhasin, Neha Goswami, Nehal Belgamwar, Nicolas Anastassacos, Nicolas Bergeron, Nidhi Jain, Nihal Jain, Niharika Chopparapu, Nik Xu, Nikko Strom, Nikolaos Malandrakis, Nimisha Mishra, Ninad Parkhi, Ninareh Mehrabi, Nishita Sant, Nishtha Gupta, Nitesh Sekhar, Nithin Rajeev, Nithish Raja Chidambaram, Nitish Dhar, Noor Bhagwagar, Noy Konforty, Omar Babu, Omid Razavi, Orchid Majumder, Osama Dar, Oscar Hsu, Pablo Kvitca, Pallavi Pandey, Parker Seegmiller, Patrick Lange, Paul Ferraro, Payal Motwani, Pegah Kharazmi, Pei Wang, Pengfei Liu, Peter Bradtke, Peter Götz, Peter Zhou, Pichao Wang, Piotr Poskart, Pooja Sonawane, Pradeep Natarajan, Pradyun Ramadorai, Pralam Shah, Prasad Nirantar, Prasanthi Chavali, Prashan Wanigasekara, Prashant Saraf, Prashun Dey, Pratyush Pant, Prerak Pradhan, Preyaa Patel, Priyanka Dadlani, Prudhvee Narasimha Sadha, Qi Dong, Qian Hu, Qiaozi, Gao, Qing Liu, Quinn Lam, Quynh Do, R. Manmatha, Rachel Willis, Rafael Liu, Rafal Ellert, Rafal Kalinski, Rafi Al Attrach, Ragha Prasad, Ragini Prasad, Raguvir Kunani, Rahul Gupta, Rahul Sharma, Rahul Tewari, Rajaganesh Baskaran, Rajan Singh, Rajiv Gupta, Rajiv Reddy, Rajshekhar Das, Rakesh Chada, Rakesh Vaideeswaran Mahesh, Ram Chandrasekaran, Ramesh Nallapati, Ran Xue, Rashmi Gangadharaiah, Ravi Rachakonda, Renxian Zhang, Rexhina Blloshmi, Rishabh Agrawal, Robert Enyedi, Robert Lowe, Robik Shrestha, Robinson Piramuthu, Rohail Asad, Rohan Khanna, Rohan Mukherjee, Rohit Mittal, Rohit Prasad, Rohith Mysore Vijaya Kumar, Ron Diamant, Ruchita Gupta, Ruiwen Li, Ruoying Li, Rushabh Fegade, Ruxu Zhang, Ryan Arbow, Ryan Chen, Ryan Gabbard, Ryan Hoium, Ryan King, Sabarishkumar Iyer, Sachal Malick, Sahar Movaghati, Sai Balakavi, Sai Jakka, Sai Kashyap Paruvelli, Sai Muralidhar Jayanthi, Saicharan Shriram Mujumdar, Sainyam Kapoor, Sajjad Beygi, Saket Dingliwal, Saleh Soltan, Sam Ricklin, Sam Tucker, Sameer Sinha, Samridhi Choudhary, Samson Tan, Samuel Broscheit, Samuel Schulter, Sanchit Agarwal, Sandeep Atluri, Sander Valstar, Sanjana Shankar, Sanyukta Sanyukta, Sarthak Khanna, Sarvpriye Khetrapal, Satish Janakiraman, Saumil

Shah, Saurabh Akolkar, Saurabh Giri, Saurabh Khandelwal, Saurabh Pawar, Saurabh Sahu, Sean Huang, Sejun Ra, Senthilkumar Gopal, Sergei Dobroshinsky, Shadi Saba, Shamik Roy, Shamit Lal, Shankar Ananthakrishnan, Sharon Li, Shashwat Srijan, Shekhar Bhide, Sheng Long Tang, Sheng Zha, Shereen Oraby, Sherif Mostafa, Shiqi Li, Shishir Bharathi, Shivam Prakash, Shiyuan Huang, Shreya Yembarwar, Shreyas Pansare, Shreyas Subramanian, Shrijeet Joshi, Shuai Liu, Shuai Tang, Shubham Chandak, Shubham Garg, Shubham Katiyar, Shubham Mehta, Shubham Srivastav, Shuo Yang, Siddalingesha D S, Siddharth Choudhary, Siddharth Singh Senger, Simon Babb, Sina Moeini, Siqi Deng, Siva Loganathan, Slawomir Domagala, Sneha Narkar, Sneha Wadhwa, Songyang Zhang, Songyao Jiang, Sony Trenous, Soumajyoti Sarkar, Soumya Saha, Sourabh Reddy, Sourav Dokania, Spurthideepika Sandiri, Spyros Matsoukas, Sravan Bodapati, Sri Harsha Reddy Wdaru, Sridevi Yagati Venkateshdatta, Srikanth Ronanki, Srinivasan R Veeravanallur, Sriram Venkatapathy, Sriramprabhu Sankaraguru, Sruthi Gorantla, Sruthi Karuturi, Stefan Schroedl, Subendhu Rongali, Subhasis Kundu, Suhaila Shakiah, Sukriti Tiwari, Sumit Bharti, Sumita Sami, Sumith Mathew, Sunny Yu, Sunwoo Kim, Suraj Bajirao Malode, Susana Cumplido Riel, Swapnil Palod, Swastik Roy, Syed Furqhan, Tagyoung Chung, Takuma Yoshitani, Taojiannan Yang, Tejaswi Chillakura, Tejwant Bajwa, Temi Lajumoke, Thanh Tran, Thomas Gueudre, Thomas Jung, Tianhui Li, Tim Seemman, Timothy Leffel, Tingting Xiang, Tirth Patel, Tobias Domhan, Tobias Falke, Toby Guo, Tom Li, Tomasz Horszczaruk, Tomasz Jedynak, Tushar Kulkarni, Tyst Marin, Tytus Metrycki, Tzu-Yen Wang, Umang Jain, Upendra Singh, Utkarsh Chirimar, Vaibhav Gupta, Vanshil Shah, Varad Deshpande, Varad Gunjal, Varsha Srikeshava, Varsha Vivek, Varun Bharadwaj, Varun Gangal, Varun Kumar, Venkatesh Elango, Vicente Ordonez, Victor Soto, Vignesh Radhakrishnan, Vihang Patel, Vikram Singh, Vinay Varma Kolanuvada, Vinayshekhar Bannihatti Kumar, Vincent Auvray, Vincent Cartillier, Vincent Ponzo, Violet Peng, Vishal Khandelwal, Vishal Naik, Vishvesh Sahasrabudhe, Vitaliy Korolev, Vivek Gokuladas, Vivek Madan, Vivek Subramanian, Volkan Cevher, Vrinda Gupta, Wael Hamza, Wei Zhang, Weitong Ruan, Weiwei Cheng, Wen Zhang, Wenbo Zhao, Wenyan Yao, Wenzhuo Ouyang, Wesley Dashner, William Campbell, William Lin, Willian Martin, Wyatt Pearson, Xiang Jiang, Xiangxing Lu, Xiangyang Shi, Xianwen Peng, Xiaofeng Gao, Xiaoge Jiang, Xiaohan Fei, Xiaohui Wang, Xiaozhou Joey Zhou, Xin Feng, Xinyan Zhao, Xinyao Wang, Xinyu Li, Xu Zhang, Xuan Wang, Xuandi Fu, Xueling Yuan, Xuning Wang, Yadunandana Rao, Yair Tavizon, Yan Rossiytsev, Yanbei Chen, Yang Liu, Yang Zou, Yangsook Park, Yannick Versley, Yanyan Zhang, Yash Patel, Yen-Cheng Lu, Yi Pan, Yi-Hsiang, Lai, Yichen Hu, Yida Wang, Yiheng Zhou, Yilin Xiang, Ying Shi, Ying Wang, Yishai Galatzer, Yongxin Wang, Yorick Shen, Yuchen Sun, Yudi Purwatama, Yue, Wu, Yue Gu, Yuechun Wang, Yujun Zeng, Yuncong Chen, Yunke Zhou, Yusheng Xie, Yvon Guy, Zbigniew Ambrozinski, Zhaowei Cai, Zhen Zhang, Zheng Wang, Zhenghui Jin, Zhewei Zhao, Zhiheng Li, Zhiheng Luo, Zhikang Zhang, Zhilin Fang, Zhiqi Bu, Zhiyuan Wang, Zhizhong Li, Zijian Wang, Zimeng, Qiu, and Zishi Li. The amazon nova family of models: Technical report and model card, 2025. URL https://arxiv.org/abs/2506.12103.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke,

Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. Openscholar: Synthesizing scientific literature with retrieval-augmented lms, 2024. URL https://arxiv.org/abs/2411.14199.

Yucheng Cai, Si Chen, Yuxuan Wu, Yi Huang, Junlan Feng, and Zhijian Ou. The 2nd futuredial challenge: Dialog systems with retrieval augmented generation (futuredial-rag), 2024. URL https://arxiv.org/abs/2405.13084.

Tianyu Cao, Neel Bhandari, Akhila Yerukola, Akari Asai, and Maarten Sap. Out of style: RAG's fragility to linguistic variation. In Vera Demberg, Kentaro Inui, and Lluís Marquez, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 280–318, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.13. URL https://aclanthology.org/2026.eacl-long.13/.

Qi Chen, Xiubo Geng, Corby Rosset, Carolyn Buractaon, Jingwen Lu, Tao Shen, Kun Zhou, Chenyan Xiong, Yeyun Gong, Paul Bennett, Nick Craswell, Xing Xie, Fan Yang, Bryan Tower, Nikhil Rao, Anlei Dong, Wenqi Jiang, Zheng Liu, Mingqin Li, Chuanjie Liu, Zengzhong Li, Rangan Majumder, Jennifer Neville, Andy Oakley, Knut Magne Risvik, Harsha Vardhan Simhadri, Manik Varma, Yujing Wang, Linjun Yang, Mao Yang, and Ce Zhang. Ms marco web search: A large-scale information-rich web dataset with millions of real click labels. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 292–301. ACM, May 2024. doi: 10.1145/3589335.3648327. URL http://dx.doi.org/10.1145/3589335.3648327.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

Chanyeol Choi, Jy-Yong Sohn, Subeen Pang Yongjae Lee, Jaeseon Ha, Hoyeon Ryoo, Yongjin Kim, Hojun Choi, and Jihoon Kwon. Financerag challenge., 2024. URL https://kaggle.com/competitions/icaif-24-finance-rag-challenge.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, and Qifeng Chen. Animate-a-story: Storytelling with retrieval-augmented video generation, 2023. URL https://arxiv.org/abs/2307.06940.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023. URL https://arxiv.org/abs/2311.17982.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=HfKOIPCvsv.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models, 2024. URL https://arxiv.org/abs/2310.11440.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, Colin Raffel, Adam Roberts, Tom Kwiatkowski, Patrick Lewis, Yuxiang Wu, Heinrich Küttler, Linqing Liu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Sohee Yang, Minjoon Seo, Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Edouard Grave, Ikuya Yamada, Sonse Shimaoka, Masatoshi Suzuki, Shumpei Miyawaki, Shun Sato, Ryo Takahashi, Jun Suzuki, Martin Fajcik, Martin Docekal, Karel Ondrej, Pavel Smrz, Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Barlas Oguz, Xilun

Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Wen-tau Yih. Neurips 2020 efficientqa competition: Systems, analyses and lessons learned. In Hugo Jair Escalante and Katja Hofmann, editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 86–111. PMLR, 06–12 Dec 2021. URL https://proceedings.mlr.press/v133/min21a.html.

Cattalyya Nuengsigkapian. Hifi-rag: Hierarchical content filtering and two-pass generation for open-domain rag, 2025. URL https://arxiv.org/abs/2512.22442.

Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. Clueweb22: 10 billion web documents with visual and semantic information, 2022. URL https://arxiv.org/abs/2211.15848.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. Kilt: a benchmark for knowledge intensive language tasks, 2021. URL https://arxiv.org/abs/2009.02252.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. doi: 10.1162/tacl_a_00605. URL https://aclanthology.org/2023.tacl-1.75/.

Kun Ran, Marwah Alaofi, Danula Hettiachchi, Chenglong Ma, Khoi Nguyen Dinh Anh, Khoi Vo Nguyen, Sachin Pathiyan Cherumanal, Lida Rashidi, Falk Scholer, Damiano Spina, Shuoqi Sun, and Oleg Zendel. Rmit-adm+s at the mmu-rag neurips 2025 competition, 2026. URL https://arxiv.org/abs/2602.20735.

Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.403. URL https://aclanthology.org/2024.acl-long.403/.

Shuting Wang, Xin Yu, Mang Wang, Weipeng Chen, Yutao Zhu, and Zhicheng Dou. RichRAG: Crafting rich responses for multi-faceted queries in retrieval-augmented generation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11317–11333, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.750/.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2024. URL https://arxiv.org/abs/2307.06942.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Ikuya Yamada, Wataru Ikeda, Ko Yoshida, Mengyu Ye, Hinata Sugimoto, Masatoshi Suzuki, Hisanori Ozaki, and Jun Suzuki. An open and reproducible deep research agent for long-form question answering, 2025. URL https://arxiv.org/abs/2512.13059.

Diji Yang, Linda Zeng, Jinmeng Rao, and Yi Zhang. Knowing you don't know: Learning when to continue search in multi-round rag through self-practicing. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1305–1315, 2025.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. Crag - comprehensive rag benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 10470–10490. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1435d2d0fca85a84d83ddcb754f58c29-Paper-Datasets_and_Benchmarks_Track.pdf.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL https://arxiv.org/abs/1809.09600.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net, 2020.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2023.

Appendix A. Data Curation

This section details the curation of the queries used for both text-to-text and text-to-video tracks, as well as the curation of reference answers for the text-to-text track.

Table 7: Summary of data used in the competition.

Source	Use Case	Track(s)	Size	Access & Licensing
ClueWeb22-B-EN ClueWeb22-A-EN	Retrieval corpora	Text-to-text Text-to-video	87M+ docs 800M+ docs	Index hosted via API (free), licensed and distributed by CMU[1]
MS MARCO Web Search	Static queries (knowledge-intense)	Text-to-text	1,000 queries	MIT License
Chatbot Arena Conversations	Static queries (video-intent)	Text-to-video	300 prompts	CC BY 4.0 license
RAG-Arena Live Queries	Live dynamic eval	Text-to-text Text-to-video	-	Collected from live users

Web-scale Corpora. We host API access to ClueWeb22-B-EN and ClueWeb22-A-EN (Overwijk et al., 2022) (87M+ and 800M documents) as our background corpus for web-scale retrieval settings.[2] Further, participants were allowed to use search APIs (e.g., SERP[3]) as this reflects current common practices in the field.

Static Queries and References. Static queries are used for offline evaluation of submitted systems. We release validation queries for development, but keep test queries private for final ranking. For the (A) text-to-text track, we curate a set of 1,000 knowledge-intensive queries from MS Marco Web Search (Chen et al., 2024) and Chatbot Arena Conversations (Zheng et al., 2023). Unlike conventional QA datasets that focus on concise factual answers (Kwiatkowski et al., 2019; Yang et al., 2018), knowledge-intensive queries require nuanced, explanation-based responses. We construct our dataset by using an LLM to determine whether queries are open-ended and/or explanatory based on their information-seeking intent. We then manually review and select 1,000 queries to ensure domain diversity and clarity to form a high-quality set of knowledge-intensive prompts (e.g. *"How does urbanization impact the environment?"*).

We construct reference answers by prompting an LLM to generate responses based on the top-clicked documents and user-preferred model responses. These model-generated responses are then manually reviewed and validated for relevance, coherence, and factual accuracy to ensure high-quality responses, while remaining grounded to documents users actually engaged with. Reference answers are used for reference-based evaluations (Section 3), and are not released to participants.

2. The license to distribute the data via API-access for the competition has been approved dataset authors, one of whom is an organizer of the competition. Participants can also gain access to the full corpus by signing a data license agreement: http://lemurproject.org/clueweb22/obtain.php
3. https://serpapi.com

For the text-to-video track, a set of 300 video query prompts is further filtered from the queries curated above. An LLM is used to score queries based on video intent, identifying queries for their need for visual explanation, including demonstrations, dynamic processes, and spatial reasoning (see Appendix C.) These queries are then manually reviewed for clarity and diversity across domains. Further, we validate video intent by checking whether video results appear among the top results in commercial search engines (e.g., Google), indicating that users typically expect visual content.

Appendix B. Biography of all team members

Luo Qi Chan (luoqic@andrew.cmu.edu), Carnegie Mellon University
Luo Qi is a research engineer at DSO National Laboratories currently pursuing her Masters' degree at Carnegie Mellon University. Her previous work include question-answering for industry applications, and learning representations for efficient retrieval. She is part of the team that developed the video evaluation framework for one of the competition tracks.

Tevin Wang (tevinw@andrew.cmu.edu), Carnegie Mellon University
Tevin is an undergraduate student in Computer Science at Carnegie Mellon. He previously developed RAGViz, a RAG diagnostic and visualization platform presented at EMNLP 2024 as a system demonstration. His focus is on designing RAG-Arena, the competition's live evaluation system.

Shuting Wang (shutingw@andrew.cmu.edu), Carnegie Mellon University
Shuting is a visiting scholar at the Language Technologies Institute, Carnegie Mellon University. Her research interests include information retrieval, large language models for next-generation IR, and exploring multimodal information generation. She is part of the team that developed the video evaluation framework for one of the competition tracks.

Zhihan Zhang (zhihanz@andrew.cmu.edu), Carnegie Mellon University
Zhihan is a system software engineer in Professor Chenyan Xiong's research lab at the Language Technologies Institute, Carnegie Mellon University. His primary responsibilities include designing, implementing, and testing full-stack and large-scale software systems to support research in interactive deep learning, large language models, and retrieval systems. He also manages the development and maintenance of web platforms for lab projects.

Alfredo Gomez (alfredo3@andrew.cmu.edu), Carnegie Mellon University
Alfredo is a PhD student advised by Prof. Mona Diab and Fernando Diaz. His research focuses on the evaluation and trustworthiness of NLP systems. He is part of the coordination team.

Prahaladh Chandrahasan (prahalac@andrew.cmu.edu), Carnegie Mellon University
Prahaladh is a graduate student in the Software Societal Systems department at the School of Computer Science at CMU. He is currently working as a Research Software Engineer under Mona Diab, with previous industry stints at Bank of America and RedHat. His primary focus is on live evaluation.

Lan Yan (lany2@andrew.cmu.edu), Carnegie Mellon University
Lan is a second-year master student majoring in Artificial Intelligence at Carnegie Mellon University. Her previous research works focused on LLM tool use and multimodal embedding models. She worked on developing text-to-video generation baselines for one of the competition tracks.

Andy Tang (andyt@andrew.cmu.edu), Carnegie Mellon University
Andy is a Master's student in the Computational Data Science program at Carnegie Mellon University, advised by Prof. Chenyan Xiong. His research interests lie in natural language processing and deep learning, building on prior experience as a full-stack software engineer.

Zimeng (Chris) Qiu (zimengqi@amazon.com), Amazon
Chris is a Senior Applied Scientist at Amazon AGI where he has led several critical science workstreams on building foundation models. His research focuses on building accurate and reliable retrieval-augmented generation (RAG) systems, improving foundation models core capabilities (e.g., coding), etc. Prior to joining Amazon AGI, he worked as an Applied Scientist at Alexa AI where he focused on information retrieval and ranking research.

Morteza Ziyadi (mziyadi@amazon.com), Amazon
Morteza Ziyadi is an Applied Science Manager at Amazon AGI, where he leads several projects focused on large-scale evaluations of foundation models. Before joining Amazon AGI, he spent four years at Microsoft Cloud and AI, where he led projects focused on developing natural language-to-code generation models for various products. He has also served as an adjunct faculty member at Northeastern University. He earned his Ph.D. from the University of Southern California (USC) in 2017 and has since been actively involved as workshop organizer and reviewer for numerous NLP, CV, and machine learning.

Sherry Wu (sherryw@andrew.cmu.edu), Carnegie Mellon University
Sherry is an assistant professor at the Human-Computer Interaction Institute, Carnegie Mellon University. Her primary research investigates how humans (AI experts, lay users, domain experts) interact with (debug, audit, and collaborate) AI systems. Sherry has organized two workshops at NLP and HCI conferences: Shared Stories and Lessons Learned workshop at EMNLP 2022 and Trust and Reliance in AI-Human Teams at CHI 2022 and 2023. She has given two well-received tutorials relevant to Human-AI Interaction, one at EMNLP 2023 on Designing, Learning from, and Evaluating Human-AI Interactions, and another one on Human-AI Interactions in the Era of LLMs at NAACL 2024.

Mona Diab (mdiab@andrew.cmu.edu), Carnegie Mellon University
Dr. Mona T. Diab is the Director of the Language Technologies Institute (LTI) at Carnegie Mellon University (CMU) and a Full Professor in the School of Computer Science as well as (by courtesy) a Full Professor in the CMU Heinz School of Information Management. With over 21,000 Google Scholar citations and an H-index of 59, Dr. Diab is a globally recognized expert in computational linguistics and natural language processing (NLP). Her career spans academia, industry, and international leadership roles, including a tenured full professorship at George Washington University, Amazon AWS, and Meta, where she served as Lead Responsible AI Research Scientist and Technical Lead. An ACL Fellow, she has received numerous accolades, including the King Salman Global Arabic Academy award (2023) for her contributions to Arabic language technologies and recognition among the Top Global AI Scientists of Arab Descent by MIT Technology Review.

Dr. Diab's research focuses on advancing trustworthy NLP and responsible AI systems, particularly in low-resource and multilingual contexts. Her work includes innovations in controllable natural language generation, cross-lingual processing, computational social science, and health analytics. She has spearheaded high-impact initiatives, such as democratizing access to scientific content through machine translation and addressing bias and compliance in AI systems. A pioneer in Arabic NLP, Dr. Diab has developed founda-

tional resources and tools that bridge linguistic, social, and computational disciplines. She currently directs the R3LIT Lab (pronounced "relit") at CMU.

As an active member of the international research community, Dr. Diab has co-founded initiatives like the *SEM Conference and the Computational Approaches to Linguistic Code-Switching workshops. She serves on editorial boards of leading journals and advisory committees for global AI governance. With a steadfast commitment to mentorship, she continues to shape the next generation of computational linguists while fostering innovation at the intersection of AI and societal impact.

Akari Asai (akari@cs.washington.edu), University of Washington
Akari is a Ph.D. candidate in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. Her research focuses on overcoming the limitations of large language models (LMs) by developing advanced systems, such as Retrieval-Augmented LMs, and applying them to real-world challenges, including scientific research and underrepresented languages. Her contributions have been widely recognized, earning multiple paper awards at top NLP and ML conferences, the IBM Global Fellowship, and industry grants. She was also named an EECS Rising Star (2022) and one of MIT Technology Review's Innovators Under 35 Japan. Her work has been featured in outlets such as Forbes and MIT Technology Review. Beyond her research, Akari actively contributes to the NLP and ML communities as a co-organizer of high-impact tutorials and workshops, including the first tutorial on Retrieval-Augmented LMs at ACL 2023, as well as workshops on Multilingual Information Access (MIA; NAACL 2022) and Knowledge-Augmented NLP (NAACL 2025) and led the first shared task on Multilingual Retrieval-Augmented LMs at NAACL 2022 MIA.

Chenyan Xiong (cx@andrew.cmu.edu), Carnegie Mellon University
Chenyan is an Associate Professor at Language Technologies Institute (LTI), Carnegie Mellon University (CMU). His research interests cover various aspects of foundation and large language models (LLMs), including obtaining better and new capabilities from foundation models at reduced training cost, building next-gen information systems with LLMs such as RAG and DeepResearch. Before joining CMU as a faculty, Chenyan was at Microsoft Research from 2018 to 2023, and worked on conversational search, dense retrieval, healthcare AI, and large scale pretraining. Chenyan is the co-creator of TREC Conversational Search Track, ClueWeb22 corpus, and MSMARCO-Web benchmark. Chenyan has published around 100 peer reviewed papers on top IR, NLP, and ML venues, with best paper awards, and early career awards at SIGIR.

Appendix C. Annotation Prompt for Query Labeling

The following prompt was used to identify queries where video responses provide added value compared to text-only responses:

```
You are required to label web-search queries for how well a video will value-add
to the response of this query.
Label 2 represents a video will value-add to the answer of the provided query,
following the scenarios given below:
  2A. Query is an explicit video request.
  2B. Query requires skill demonstration, such as cooking and origami folding.
      Text-only answers are limited in instructional ability in this scenario.
  2C. Knowledge-intensive query related to concepts where visual aids clarify
      relationships, processes better than static text.
      Example: \how do tectonic plates move to create earthquakes".
  2D. Query related to concepts that require spatial and/or imagery-based
      explanations. Example: "why does my knee hurt".
  2E. Query that asks to show the process of some change over time.
      Example: "How has China changed over time".

Label 1 represents that a text or video answer has the same effect, following:
  1A. Query that asks for long-form information.
      Example: \Tell me the history of the Vatican".

Label 0 represents a video is unnecessary, and text is sufficient, following:
  0A. Simple arithmetic question.
  0B. Explicit text generation requests (e.g., \write a report...").
  0C. Factual requests asking for a piece of information, often with \what" or \when".
  0D. Query asking for a simple explanation, typically starting with "why".

Please label the given query as 0 (text sufficient),
1 (text or video equally effective), or 2 (video adds value).
Also include your reason and category (e.g., 2A, 1B) when applicable.
Respond in JSON format without new lines, using the keys
'label', 'reason', and 'category'.

Query: {query}
```

Appendix D. Video Evaluation Rubric

We provide here the full rubric used for evaluating generated videos. Each dimension is rated on a five-point Likert scale, with 1 indicating the lowest quality and 5 indicating the highest quality.

Visual Quality

[**Visual quality**] "The visual quality (sharpness, resolution) was good enough to see details clearly."
1 = Very low quality – Blurry, pixelated, hard to watch.
2 = Somewhat low quality – Somewhat visible, but poor resolution.
3 = Acceptable – Not great, but I could see most of what I needed.
4 = High quality – Just a tiny bit of room for improvement.
5 = Perfect – Crystal clear, very sharp visuals.

[**Visual realism**] "The video looks realistic — the people, objects, and movements appear natural and believable."
1 = Very fake – Obvious distortions like warped hands, jerky motion, or uncanny faces.
2 = Fake-looking – Multiple elements feel off or computer-generated.
3 = Somewhat realistic – Mostly fine, but with one or two noticeably odd visuals.
4 = Realistic – Everything looks mostly natural.
5 = Highly realistic – Nothing looks off; the visuals are convincingly human and real.

[**Temporal continuity**] "The video makes sense in terms of time and continuity — it doesn't feel like it jumps abruptly or unnaturally."
1 = Very jarring – Abrupt cuts or time skips that break the illusion of continuity.
2 = Somewhat jumpy – Some confusing or unrealistic jumps.
3 = Neutral – Not smooth, but not distracting either.
4 = Smooth – Transitions are mostly natural and continuous.
5 = Seamless – Very natural and realistic time progression.

Utility for Instruction

[**Relevance**] "The video is about the task I asked for in my instruction."
1 = Completely wrong – Totally unrelated task.
2 = Mostly wrong – Matched some keywords, but clearly not the task.
3 = Somewhat correct – Related category, but not quite right.
4 = Mostly correct – Very close, with minor errors.
5 = Perfect match – Fully aligns with the request.

[**Precision / Focus**] "The video stayed focused on the task without unnecessary content."
1 = Very unfocused – Lots of unrelated or distracting content.

2 = Somewhat unfocused – Some extra material that isn't useful.
3 = Mostly focused – A little fluff, but nothing major.
4 = Well-focused – Almost everything contributes to the goal.
5 = Highly precise – No wasted time, everything is relevant.

[**Completeness of action**] "The main action (e.g., slicing, assembling) actually happens on screen."
1 = No action shown – Only object or finished task.
2 = Very little action – Movement, but task is skipped.
3 = Some action – Partial or unclear execution.
4 = Mostly executed – The action is shown but could be clearer.
5 = Fully executed – The full action is clearly demonstrated.

[**Usefulness**] "I feel confident I can correctly do the task after watching this video."
1 = Not at all – Would need to search elsewhere.
2 = Slightly confident – Might attempt but likely fail.
3 = Somewhat confident – Could try, but may need extra help.
4 = Mostly confident – Can follow along with expected success.
5 = Completely confident – Ready to perform the task immediately.

Appendix E. Human Evaluation Breakdown for Text-to-Video Track

Table 8 and 9 shows the break down of average human-evaluated likert ratings across seven categories. The highest possible rating is 5, with lowest possible rating of 1. For more details, refer to Appendix D for the video evaluation rubrics.

Table 8: Human evaluation (visual quality metrics) for the Text-to-Video track.

Team	**Vis. Qual.**	**Vis. Real.**	**Trans. Real.**
Nova-Reel (Baseline)	3.92	3.02	3.60
`DeepVideoResearcher`	3.88	3.34	3.14

Table 9: Human evaluation (relevance and utility metrics) for the Text-to-Video track.

Team	**Relevance**	**Precision**	**Utility**	**Step Qual.**
Nova-Reel (Baseline)	2.60	2.94	2.44	2.12
`DeepVideoResearcher`	2.94	3.22	3.20	2.70

Appendix F. RAG-Arena Prototype UI

An early prototype UI can be found below:

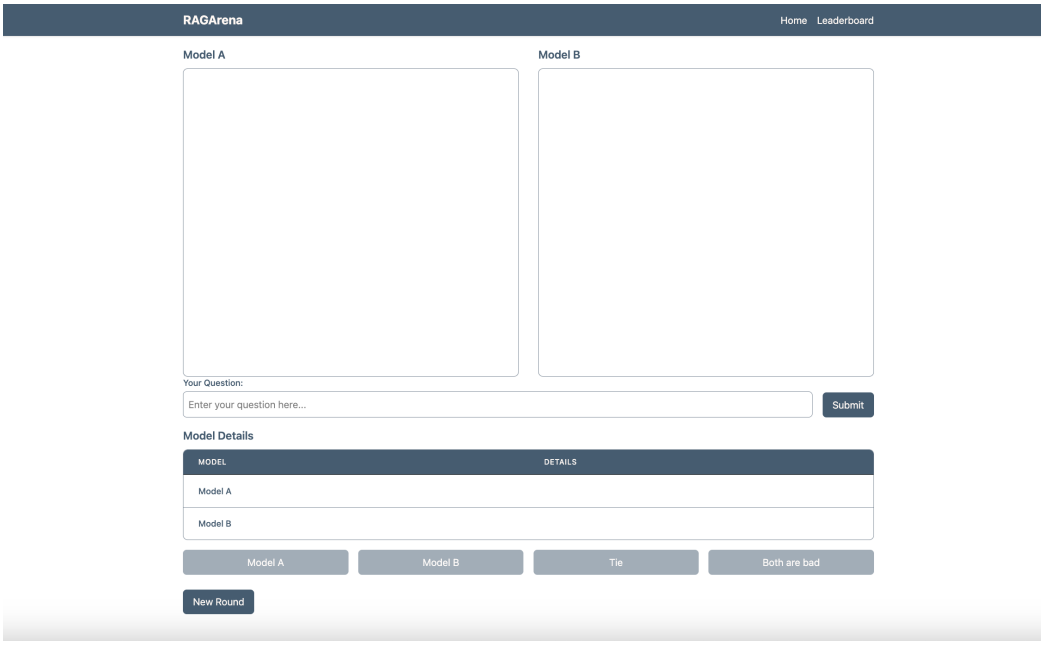


Figure 2: Early prototype UI of RAG-Arena.

Appendix G. Onboarding

G.1. Participant Registration

Participants can register for either of the two tracks (Text-to-Text, Text-to-Video) or may choose to only upload their submissions for the validation set by filling out the following form: https://forms.gle/YDEnjV4PXWnZdfYG8. Participants are required to fill in the team name, team correspondent's name followed by the names and email addresses of their teammates. The participants then choose the track in which they want to participate. Teams are encouraged to participate in more than one track, and if they choose to do so, they are required to fill out a fresh registration form for each track.

Validation file uploads Teams that choose to upload outputs only for the validation set are assigned a unique team ID and a Google folder to upload their generated outputs. Participants registered for the text-to-text track are given a folder at a shared drive managed by the organizers. Participants should then inform the organizers about the submission by filling out the submission form.

G.2. Text-to-Text Track

Participants registered for the Text-to-Text track will have a dedicated container registry in the SCS-LTI AWS account. Each team will also be given AWS-CLI access to the container

registry to through access keys that strictly only allow the teams to access their assigned container registry. The JSON policy governing the participant key access is shown below :

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "ECRLogin",
            "Effect": "Allow",
            "Action": "ecr:GetAuthorizationToken",
            "Resource": "*"
        },
        {
            "Sid": "PushPullOnlyThisRepo",
            "Effect": "Allow",
            "Action": [
                "ecr:BatchCheckLayerAvailability",
                "ecr:CompleteLayerUpload",
                "ecr:InitiateLayerUpload",
                "ecr:PutImage",
                "ecr:UploadLayerPart",
                "ecr:BatchGetImage",
                "ecr:GetDownloadUrlForLayer",
                "ecr:DescribeImages",
                "ecr:ListImages"
            ],
            "Resource": "THIS TEAM'S ECR_ARN"
        }
    ]
}
```

Teams are expected to push their docker images in a format mentioned in the competition website subject to given conditions.

G.2.1. Text-to-Video Track

Participants in the text-to-video track must containerize their video RAG systems and store all query outputs as **output.mp4** files in their assigned S3 bucket, then return the S3 object URI for each submission. Following the same infrastructure model as the text-to-text track, participants will receive a dedicated container registry, access keys, and an S3 bucket, with the same set of keys providing access to both the container registry and S3 storage. The JSON policy governing the participant key access is shown below :

```
{
    "Version": "2012-10-17",
```

```
    "Statement": [
        {
            "Sid": "VisualEditor0",
            "Effect": "Allow",
            "Action": [
                "s3:DeleteObjectVersion",
                "s3:GetObjectVersionTagging",
                "s3:RestoreObject",
                "s3:ListBucket",
                "ecr:UploadLayerPart",
                "ecr:ListImages",
                "s3:GetObjectVersionAttributes",
                "s3:ReplicateObject",
                "s3:GetObjectVersionTorrent",
                "s3:GetObjectAcl",
                "s3:GetObjectVersionAcl",
                "ecr:CompleteLayerUpload",
                "s3:GetObjectTagging",
                "s3:DeleteObject",
                "s3:PutObjectAcl",
                "ecr:BatchCheckLayerAvailability",
                "s3:GetObjectRetention",
                "s3:GetBucketWebsite",
                "ecr:GetDownloadUrlForLayer",
                "s3:GetObjectAttributes",
                "s3:GetObjectLegalHold",
                "ecr:PutImage",
                "s3:PutObject",
                "s3:GetObject",
                "s3:GetObjectTorrent",
                "s3:PutObjectRetention",
                "ecr:BatchGetImage",
                "ecr:DescribeImages",
                "s3:GetObjectVersionForReplication",
                "ecr:InitiateLayerUpload",
                "s3:GetObjectVersion"
            ],
            "Resource": [
                "s3 Object ARN",
                "ecr ARN"
            ]
        }
    ]
}
```

A sample snapshot of a generated video stored in an S3 bucket is available at the competition website

Appendix H. Partipant Systems

H.1. Team: Cattalyya

System Summary. The HiFi-RAG methodology (Nuengsigkapian, 2025) prioritizes precision in the context window by replacing standard vector-similarity search with a five-stage hierarchical filtering pipeline that leverages the speed of Gemini 2.5 Flash and the reasoning of Gemini 2.5 Pro. The process begins with query formulation to transform verbose user questions into optimized search queries, followed by URL filtering to discard irrelevant domains before scraping. Instead of treating content as flat text, the system uses hierarchical content parsing to structure web and Reddit data into trees (e.g., nested comments or header-grouped text). To ensure high-quality context, an LLM-based reranker filters these sections based on titles and snippets , feeding the results into a two-pass generation phase where the answer is drafted and then refined for style , concluding with a post-hoc citation verification step to guarantee accurate attribution.

Lessons & Challenges. The team observed that prompt engineering yielded the highest impact on performance, followed by content and URL filtering for content quality. Agentic workflows proved significantly slower and more expensive than deterministic pipelines, without necessarily improving performance on known tasks. Furthermore, we found that engineering execution, specifically the success rates of scrapers and web parsers, was just as critical to overall quality as the research methodology. The team also noted that improving upon the base model was difficult, as optimized prompts often provided sufficient answers from the given validation set. To strictly evaluate retrieval, I prompted Gemini 3 Pro to create the "Test2025" dataset (targeting events after the January 2025 knowledge cutoff); on this data, baseline performance dropped significantly, allowing HiFi-RAG to demonstrate a 57.4% ROUGE-L improvement by relying on retrieved context.

H.2. Team: DeepVideoResearcher

System Summary. DeepVideoResearcher is a multi-round, retrieval-augmented video generation system that treats video synthesis as a research problem rather than a one-shot generation step. The system decomposes user prompts into structured research questions, performs iterative web-based retrieval to fill gaps in visual and factual knowledge, and synthesizes this information into a verified visual context brief. Following their prior work in text-based RAG system (SIM-RAG; Yang et al. (2025)), they deployed an LLM-based external Critic Agent to evaluate whether the retrieved evidence is sufficient, consistent, and aligned with user intent before any generation occurs. Only upon a successful verification loop does the system translate the grounded brief into a generator-ready prompt for diffusion-based video models. By inserting active research and external critique prior to generation, DeepVideoResearcher produces videos that are factually grounded, reliable, and suitable for scientific, educational, and production-grade use cases

Lessons & Challenges. One major challenge the team encountered was the high cost (both run time and compute) of video generation when using state-of-the-art diffusion-based video models. Unlike text or image generation. This constraint forced the team to design the system so that generation only occurs at the final step, and not part of the iterative

loop. As a result, they placed strong emphasis on upstream validation to ensure that research, grounding, and critique converge before invoking the video generator. This serves to minimize wasted compute on incorrect or poorly grounded contexts. Another key lesson involved refining the external critique loop. While the Critic Agent is essential for preventing hallucinated grounding, calibrating its strictness proved non-trivial. Early versions of the system suffered from over-critique, where the agent repeatedly requested additional evidence even when sufficient information was already available, slowing down the pipeline. Conversely, a lenient critic allowed subtle gaps in visual detail to pass through, leading to downstream generation errors. Achieving the right balance required iterative tuning of critique criteria and explicit definitions of what constitutes "sufficient" visual grounding for video generation. Finally, their current optimization strategy is concentrated exclusively on the alignment between the deep research output and the final input prompt for the video generator, i.e., the Critic. This alignment, enforced by the Critic Agent, has proven sufficient to achieve high-performance results by treating the video model as a fixed component. However, this creates a clear path for future advancement: while optimizing the textual bridge is effective, pushing the success of this system further will likely require moving beyond text-only optimization to include the fine-tuning or joint optimization of the video generator itself. Overall, these lessons highlight that reliability in multimodal generation depends critically on robust control flow, alignment between modular components, and the holistic optimization of multiple models.

H.3. Team: Efficient-Deep-Research

System Summary. The team proposed a deep research system that generates long-form, report-style answers (Yamada et al., 2025)[4]. To ensure reproducibility, the system is constructed using open-source Qwen 3 LLMs and a search API built upon the ClueWeb22 open web corpus. Their approach involves high-quality synthetic data generation, preference tuning, and an enhanced search component. The system is trained via Direct Preference Optimization (DPO) with LLM-as-a-Judge-based preference signals.

Lessons & Challenges. The team's experimental results show that preference tuning via DPO is highly effective for long-form deep research tasks. They successfully implemented preference tuning without prior supervised fine-tuning, as the base Qwen 3 Next models inherently follow complex instructions, including reasoning, tool use, and the generation of both citations and answers. However, they noticed that a side effect of the training is a significant increase in response length, which may not be suitable for questions that can be answered with a short response, such as factoid questions.

H.4. Team: RMIT-ADMS IR

System Summary. The R2RAG (Routing-to-RAG) system implements a dynamic retrieval-augmented generation architecture that adapts its behavior based on question complexity (Ran et al., 2026)[5]. The system relies on a question classifier that routes each input to one of two execution paths: Vanilla RAG, designed for simpler queries that can be answered with

4. Efficient-Deep-Research Github Repository: https://github.com/efficient-deep-research/efficient-deep-research
5. [6]

a single retrieval pass, and Vanilla Agent, designed for more complex queries that require iterative retrieval and evidence accumulation. The Qwen3-4B model is used for question classification and, in both paths, for query variant generation and answer generation, while Qwen3-Reranker-0.6B is used for document reranking. In the Vanilla RAG path, the system generates three query variants from the input question, retrieves up to 30 documents from ClueWeb22-A, reranks them, and uses the top-ranked documents truncated to 5,000 words to generate and stream the final answer. The Vanilla Agent maintains a record of issued queries and document summaries across iterations. A review component assesses information coverage and identifies remaining knowledge gaps to guide subsequent retrieval steps. The system terminates retrieval and produces the final answer when coverage is deemed sufficient, when the accumulated context exceeds 20,000 tokens, or after a maximum of five iterations. System development was informed by ongoing qualitative evaluation, including an extended session with 20 participants from diverse academic backgrounds who freely tested the system and provided feedback.

Lessons & Challenges. The team's experience highlights several practical lessons for deploying dynamic RAG systems under realistic constraints. First, they found that test-time scaling with small reasoning-oriented models can be effective for RAG pipelines. In their setting, recent reasoning-enabled models such as Qwen3-4B consistently outperformed larger, earlier-generation non-reasoning models (8B–10B), despite their smaller parameter count. Strong instruction following proved critical for components that require structured decisions, such as the information and coverage review module. More generally, retrieval quality remained a primary driver of answer quality: improved ranking led directly to more accurate and complete responses. For complex questions, accumulating evidence across multiple retrieval iterations, up to a budget of 20,000 tokens, provided sufficient information density for reliable answer generation. The additional prompt processing overhead was acceptable, as prefill costs remained small relative to decoding time during inference.

From a deployment perspective, the shared infrastructure imposed constraints that shaped several design choices. The team implemented the system as a single Docker image, without external services beyond the provided search API, to ensure portability and reproducibility. Supporting concurrent requests while keeping the image under 20GB and maintaining reasonable startup time required careful resource management. The team adopted vLLM without flashinfer compilation and manually tuned context length, batch size, and KV cache parameters to balance GPU memory usage between the generator and the reranker. Although Docker simplified environment isolation, integrating a pre-trained logistic regression router introduced compatibility and maintenance challenges. Replacing it with an LLM-based prompt classifier simplified the pipeline and improved robustness, while maintaining acceptable throughput. Overall, these experiences suggest that simplicity, tight integration, and careful resource tuning are key to deploying effective and efficient RAG systems in constrained evaluation settings.

H.5. Team: NightFeats

System Summary. The team implemented a context-optimized multi-agent retrieval-augmented generation (RAG) framework that decomposes text generation into three coor-

dinated phases: retrieval, curation, and composition. [7] _neurips2025 Inspired by Agentic Context Engineering (ACE), the architecture employs three specialized agents: research agent, generator agent, and writer agent operating through structured intermediate representations stored in a shared research context. The research agent performs hybrid retrieval by dynamically routing queries between Fineweb (archived web data) and Brave Search (real-time web content) based on temporal-semantic signals, providing query decomposition, semantic chunking (max-token overlap), OpenAI text-embedding-3-small embeddings, and Voyage AI reranking with temporal-semantic Xfusion (weighted 30% recency, 70% relevance). The generator agent extracts atomic facts, clusters them to detect contradictions classified by severity (minor, important, critical), and invokes bounded reconciliation cycles (maximum 2 iterations) to resolve conflicts using targeted sub-retrievals. The writer agent synthesizes curated facts into coherent, fully-cited prose with automatic citation insertion from intermediate (doc-id, format in Markdown [1], [2]) references. While their implementation leverages shared APIs (OpenAI for LLM and embeddings, Voyage AI for reranking) to focus development effort on core system architecture, the framework is built on the OpenAI Agents SDK which supports local LLM backends, and both reranking and embedding components can be replaced with self-hosted open-source alternatives, enabling fully on-premise deployment for privacy-dependent or cost-constrained environments. This structured pipeline ensures factual stability, temporal relevance, and provenance preservation while maintaining competitive semantic performance with enhanced transparency and verifiability through explicit citation traceability.

Lessons & Challenges. The team shared three challenges.

First, deploying a multi-agent RAG system in production environments requires careful balancing between response quality and latency constraints. The system's three-agent architecture featuring iterative curation cycles, external API dependencies, and comprehensive fact verification naturally introduces significant latency remarks ranging from 10-15 seconds per query in production mode compared to 6-8 seconds in development configurations. The competition requirement to handle 10-15 concurrent requests amplified this challenge, as each agent invocation consumes API quota and introduces network round-trips. We implemented several optimizations: (1) configuration modes that trade quality for speed, completed through experimentation with different GPT-5 reasoning effort levels and reduced retrieval parameters, subsequently achieving 2-3x speedups at modest quality costs, (2) parallel user expansion and document fetching to minimize sequential API calls, (3) Gunicorn with 8 workers to handle concurrent load without request queuing, and (4) extended timeouts (700s+) to accommodate the full multi-step reasoning pipeline without premature termination. Despite these efforts, the fundamental tension remains unresolved: users expect both comprehensive, verified answers and near-instantaneous responses. Thus, as different use cases demand different balances, production RAG systems must make explicit quality-latency trade-offs and expose configuration options rather than optimizing for a single operating point.

Second, a significant challenge emerged from the disconnect between standard automatic evaluation metrics (ROUGE-L, BERTScore) and the system's core design principles around transparency and verifiability. While the system maintained competitive perfor-

7. NightFeats Github Repository: https://github.com/quent1fvr/MMU_RAG

mance against strong baselines, we observed that inline citations negatively impacted automatic similarity scores. Each factual claim in the output includes explicit numeric citations ([1], [2], [3]) linked to source URLs, whereas gold referenced typical lack such citations. When the system generates "Quantum computing utilizes qubits for parallel computation [1][2]" and the reference states "Quantum computing utilizes qubits for parallel computation," the citation tokens reduce lexical overlap despite improving factual traceability. This demonstration highlights a broader evaluation challenge in RAG research: systems optimized for verifiability, provenance preservation, and user trust are penalized by metrics designed to measure fluency and semantic similarity. A paradox thus introduced underscores the concept that features improving real-world utility—explicit source attribution, structured citations, and transparency about information sources actively harm automatic evaluation scores. These instances suggest the RAG community needs evaluation frameworks that explicitly reward transparency and verifiability rather than treating them as auxiliary considerations. As a result, future benchmarks should incorporate citation quality, source reliability assessment, and factual grounding metrics alongside traditional similarity.

Last, resolving factual contradictions in retrieved evidence presents a fundamental epistemic challenge: determining which source to trust when multiple authoritative-seeming documents conflict, especially without access to ground truth. The generator agent detects contradictions by clustering facts and classifying conflicts by severity (minor, important, critical), but resolving these conflicts requires heuristics rather than definitive answers. The team implemented temporal preference (newer sources generally override older ones) combined with bounded reconciliation cycles (maximum 2 iterations with targeted sub-retrievals). However, this temporal bias introduces vulnerability to disinformation, particularly in today's information ecosystem where fake news, clickbait, and low-quality content propagate rapidly across the web. A recent fabricated story can surface higher in Brave Search results than authoritative sources simply due to recency and engagement metrics. The system partially mitigates this through source diversity (combining Fineweb's curated archive with real-time web search) and adaptive temporal weighting (50% recency, 70% relevance rather than pure temporal ranking), but fundamentally cannot distinguish genuine updates from coordinated misinformation campaigns without extensive verification. The core challenge extends beyond individual contradictions to the systemic problem of operating in adversarial information environments where misinformation campaigns deliberately exploit temporal ranking and engagement metrics. The challenge is systemic because RAG systems now operate in adversarial information environments where recency, popularity, and coordination can outweigh factual authority. As a result, resolving contradictions cannot rely on simple heuristics and instead requires mechanisms such as source authority modeling, cross-backend consistency checks, or explicit uncertainty quantification rather than forced consensus. Each of these approaches directly addresses a failure mode of adversarial retrieval, but does so at the cost of increased system complexity and additional latency.

H.6. Team: TTT-DR

System Summary. The system implements the Test-Time Diffusion Deep Researcher (TTD-DR) framework, which conceptually shifts RAG from a linear retrieve-then-generate

workflow to an iterative "denoising" process.[8] The system maintains an evolving draft report that serves as the central context anchor; in each iteration, the system uses the current draft to formulate targeted search queries, retrieves documents via the FineWeb API, and employs a neural reranker (Qwen3-Reranker) to select relevant chunks. These chunks are synthesized to "denoise" and refine the draft, progressively adding detail and correcting inaccuracies. To further enhance quality, we employ component-wise self-evolution, where intermediate steps (planning, answer synthesis) are generated, critiqued by an LLM-as-a-judge, and merged to produce superior inputs for the diffusion process.

Lessons & Challenges. A primary lesson learned was the critical trade-off between "test-time compute" and latency. The theoretical TTD-DR framework suggests deep iterative cycles (up to 20 iterations) and generating multiple diverse variants for self-evolution to maximize quality. However, in the context of the competition's evaluation timeouts, this was computationally prohibitive. They found that while the diffusion approach yields high coherence, we had to aggressively prune the hyperparameters—reducing iterations to 3 and minimizing evolution candidates—to maintain acceptable throughput. This highlighted that while increasing test-time inference scales intelligence, practical deployment requires dynamic budgeting of these iterations based on query complexity. Technically, the most significant deployment challenge was optimizing a dual-model architecture on a single GPU. We utilized vLLM to serve both the generative model (Qwen/Qwen3-4B-Instruct) and the reranker (Qwen3-Reranker-0.6B) concurrently. Integrating the reranker was particularly challenging; standard implementations were too slow for the high volume of documents retrieved from FineWeb (top-k=50). The team solved this by deploying the reranker using vLLM's score API with specific sequence-classification overrides. This allowed us to perform high-speed, batch-parallel reranking, which was essential for filtering noise from the web search results without bottlenecking the pipeline.

8. TTT-DR Github Repository: https://github.com/eamag/MMU-RAG-competition