RAG as a Content-Analysis Assistant: Auditing SDG Discourse in Online Videos

Cristian Safta
EPFL
Lausanne, Switzerland
saftakrist@gmail.com

Victor Bros
Idiap Research Institute and EPFL
Martigny, Switzerland
victor.bros@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute and EPFL
Martigny, Switzerland
gatica@idiap.ch

Abstract

Auditing how the Sustainable Development Goals (SDGs) are invoked in public communication is important for accountability, e.g., to detect superficial or strategic "SDG-washing", but the currently most accurate and interpretable form of auditing – manual content analysis – does not scale to the volume and length of contemporary online video. Long-form online video transcripts are particularly challenging: they are noisy, weakly structured, and difficult to navigate while still requiring traceable evidence for coding decisions. We study retrieval-augmented generation (RAG) as an assistant to support SDG discourse analysis, helping analysts locate relevant transcript spans and draft codebook-style codes and claim/theme summaries with quotes. We instantiate this workflow on a French SDG-focused online video corpus and evaluate a set of practical pipeline variants and their qualitative contributions. Using a compact evaluation protocol, we find that query rewriting yields the largest and most consistent end-to-end gains. The main remaining bottleneck is provenance: models often produce plausible answers while blurring the boundary between retrieved evidence and background knowledge, motivating explicit guardrails in the analytical pipeline.

CCS Concepts

• **Information systems** → **Information retrieval**; *Multimedia and multimodal retrieval*; Evaluation of retrieval results; • **Computing methodologies** → *Natural language generation.*

Keywords

retrieval-augmented generation, content analysis, SDG, online video

ACM Reference Format:
Cristian Safta, Victor Bros, and Daniel Gatica-Perez. 2026. RAG as a Content-Analysis Assistant: Auditing SDG Discourse in Online Videos. In *The 7th International Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR '26), June 16–19, 2026, Amsterdam, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3810987.3815539

1 Introduction

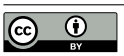
The United Nations SDGs are a widely adopted framework for sustainable development and a common reference point in institutional and public communication [24, 28]. As SDG language spreads across organizations, auditing SDG discourse – who invokes which goals, with what claims, and with what evidence – matters for accountability, including the detection of superficial or strategic references ("SDG-washing") [4, 11, 17].

Long-form, online audio-visual media channels are an increasingly important venue for this discourse, but manual content analysis does not scale. YouTube and podcast transcripts are long, noisy (ASR errors), and weakly structured, while corpora can include tens of thousands of items [13, 27]. Analysts therefore face a practical bottleneck: producing comparable coding decisions while keeping evidence inspectable and traceable at corpus scale [2, 20].

RAG is a natural candidate for this bottleneck because it conditions generation on retrieved evidence, offering more updatable and inspectable behavior than purely parametric generation [31]. However, RAG is not "plug-and-play" for analyst workflows over long-form, noisy transcripts: retrieval errors can surface irrelevant or misleading context that propagates into downstream generation [31], and the resulting outputs can blur the boundary between retrieved evidence and model background knowledge. These risks are amplified in multilingual settings, where retrieval and instruction-following can depend on the datastore language and prompt language, motivating explicit bilingual evaluation, rather than assuming English-centric results transfer unchanged to another language, e.g. a French evidence store [5].

This paper studies RAG as a pragmatic *assistant* for SDG discourse analysis in long-form video transcripts. We frame the model as an evidence-grounded drafting tool for human review: it produces structured, codebook-style coding and claim/theme extractions backed by verbatim transcript quotes, and it surfaces candidate comparator videos via retrieval to support cross-item analysis. We instantiate this approach over a French SDG-focused YouTube corpus with 13k+ videos, and we further examine an optional multimodal extension that incorporates thumbnails as auxiliary context.

Background and positioning. Prior RAG work catalogs key design choices and improvements, including pre-retrieval query transformation, retriever-side decisions such as chunking, and post-retrieval reranking and controlled generation to mitigate hallucinations [31]. Multimodal RAG extends the same principle beyond text by retrieving and integrating visual, audio, or video evidence at generation time [18]. In video-centric settings, retrieval is often performed over text-derived video representations (subtitles/ASR), while thumbnails or keyframes support analysis of visual framing [9, 12, 15, 16]. We connect this literature to analyst-oriented content analysis, where structured codes and traceable quotes are central, and semi-automation is most useful when humans retain interpretive authority [7]. Long-form media transcripts are especially challenging inputs due to noise and weak structure [1, 27], and SDG discourse analysis has largely emphasized more structured channels



ICDAR '26, Amsterdam, Netherlands

ACM ISBN 979-8-4007-2699-6/26/06
https://doi.org/10.1145/3810987.3815539

[23]; this motivates evaluating RAG under the practical constraints and failure modes that arise in transcript-based auditing.

Research questions and contributions. We pose three questions. *RQ1*: to what extent can RAG support evidence-grounded content analysis over long-form SDG video transcripts?; *RQ2*: how sensitive is the analytical system performance to RAG enhancement choices, and which improvements generalize across query types?; and *RQ3*: what are the dominant failure modes observed, and under which conditions do they occur? Our contributions are a practical RAG pipeline tailored to SDG transcript analysis (including metadata and thumbnails); a compact evaluation protocol combining LLM-as-judge comparisons across pipeline variants with held-out transcript coding, thumbnail coding, and hallucination stress tests; and empirical findings with a failure-mode taxonomy that delineate when the assistant is reliable and where it breaks.

2 Data, Tasks, and System Overview

We study retrieval-augmented generation (RAG) in a setting motivated by content analysis: analysts must efficiently locate evidence, draft structured codes, and compare discourse patterns across many long-form items.

Corpus and tasks. The corpus is a French-language YouTube dataset focused on discourse related to the 17 SDGs. It contains 13,022 videos collected in 2024-2025 using SDG-specific search queries crafted from relevant keywords for each goal. For each video, only lightweight artifacts were stored: video metadata (title, description, channel, publish date, engagement statistics), auto-generated captions (ASR transcripts), and the thumbnail image. Average video duration is *50:23* (median: *32:16*). The raw footprint is ∼1 GB (JSON + images), enabling iterative indexing and evaluation. We study three analysis actions: (i) *evidence-grounded lookup* (answer targeted questions by retrieving and quoting transcript spans); (ii) *structured coding* (draft codebook-style labels for a target video, with verbatim quotes as justification); and (iii) *cross-video comparison* (retrieve and summarize candidate comparator videos to support pattern finding).

System overview. We describe the end-to-end pipeline and the design choices we evaluate. Each corpus item is represented by (i) auto-generated captions (primary textual evidence), (ii) lightweight metadata (title, description, channel, publish date, engagement counters), and (iii) the thumbnail image.

Text indexing. To enable dense retrieval over long-form transcripts, we chunk captions into overlapping windows using a tokenizer-aware sliding strategy. We set the chunk length as large as possible within the model context (510 tokens) and use a stride of 460 tokens (50-token overlap) [30]. Each chunk is embedded with `intfloat/multilingual-e5-large` (1024-d) [29], using the model-recommended query/passage prefixes and L2-normalization. Embeddings and chunk metadata (videoId, title, statistics, etc.) are stored in an on-disk ChromaDB collection [6]. This design keeps retrieval fast and easy to iterate on.

Retrieval and retrieval-time enhancements. Given an analyst query (e.g., "What claims are made about SDG 13?" or "Does this video mention partnerships with NGOs?"), the baseline retriever encodes the query with the same embedding model and returns the top-k nearest chunks under the vector-store distance metric (L2 by default in ChromaDB) [6]. We evaluate two optional retrieval enhancements that trade additional compute for better evidence quality:

- *Query rewriting:* an LLM rewrites the user query into a retrieval-oriented form (expanding key entities, SDG terminology, and likely paraphrases) before embedding.
- *LLM reranking:* we over-retrieve a larger candidate pool and ask an LLM to score each chunk for relevance to the query under a simple rubric; the top-k reranked chunks are then passed to generation.

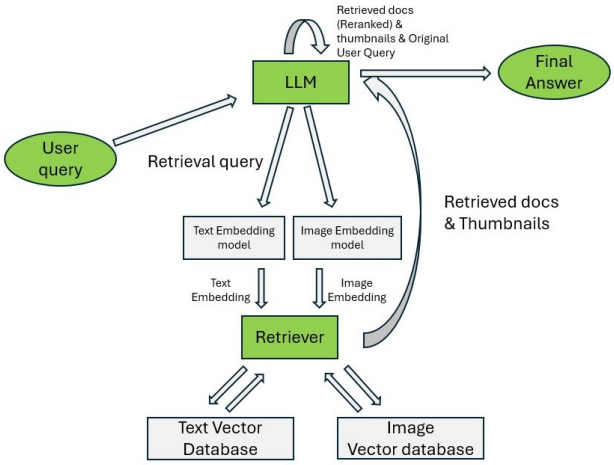


Figure 1: System architecture: optional query rewriting, text/image retrieval, reranking, and LLM generation.

Generation, structured outputs, and guardrails. The generator consumes the user query plus the retrieved chunks, and produces structured outputs. In text-only mode, we use `Llama-3.1-8B-Instruct` [10]. For multimodal variants, we use `Mistral--3.1-24B-Instruct-2503` [19]. We use a minimal RAG prompt template that provides the user question followed by the retrieved context. The model is instructed to use the context when relevant, and otherwise rely on its own reasoning or state that it does not know. As a lightweight navigation aid for long videos, we optionally attach an approximate timestamp range to each retrieved chunk by mapping its relative position in the transcript to the video duration; this is used for analyst convenience rather than as ground-truth alignment.

Thumbnail handling. We experimented with two thumbnail designs. *(A) Text-aligned thumbnails* attach the thumbnail of each video whose transcript chunks were retrieved. This guarantees that the image corresponds to the retrieved text chunk, but it cannot surface visually important videos that are missed by caption retrieval. *(B) Independent thumbnail retrieval* treats thumbnails as a first-class retrievable modality: all thumbnails are embedded with a CLIP-style encoder (`jinaai/jina-clip-v2`) [14] and searched with the query in the same image-embedding space. This surfaces visually relevant thumbnails but can decouple images from retrieved text. This trade-off suggests a natural next step: retrieve coherent (chunk, thumbnail) pairs via a joint score under a same-video constraint.

3 Evaluation Design

Overview. We evaluate the pipeline as an *analyst assistant* rather than a standalone classifier: outputs should remain inspectable (structured fields and quotes) and degrade safely when evidence is missing. We combine a small-scale benchmark for fast variant selection with targeted qualitative probes of analyst-facing behavior.

LLM-as-judge benchmark (variant selection). We construct a 15-question benchmark spanning (a) general SDG Q/A (e.g., "What is SDG 1?"), (b) detail-oriented retrieval (numbers, statistics and statements), and (c) analyst-style synthesis (claims, framing, and cross-video comparison). For each variant (top-k, query rewriting, reranking, multimodal mode, EN vs FR version), we run end-to-end RAG and score outputs with a rubric-based judge LLM across six metrics summarizing relevance/coverage, grounding to retrieved evidence, use of retrieved context, analyst usefulness, and provenance transparency (including an overall quality score). Scores are aggregated across questions to identify the strongest configuration for deeper analysis.

Held-out transcript coding. To test analyst-style coding on *unseen* items, we sample 50 videos and remove their transcript chunks from the retrieval pool. The model is given (i) the full transcript of the target video, and (ii) a *mock codebook*, and is tasked with producing a structured coding sheet.

The codebook includes a block for (a) assigning codes grounded in verbatim transcript quotes, (b) extracting the video's main claims with supporting quotes, and (c) proposing cross-video links by retrieving candidate comparator chunks from the remaining corpus.

We manually inspect all 50 outputs for query rule adherence (format, quoting, etc.) and code plausibility; for 10 videos we additionally create a small human-coded reference to characterize agreement and recurring divergences.

Multimodal thumbnail coding and hallucination stress test. For the vision-capable variant, we evaluate whether the model can apply simple visual codes to retrieved thumbnails (e.g., depicted entities, clarity score, clickbait score) without inventing content. Separately, we run a "plausible-but-unanswerable" stress test (5 queries) designed to elicit overconfident completions (e.g., corpus-wide statistics or exact quotes not present in retrieved context). We compare a minimal prompt against a guarded prompt emphasizing an explicit "insufficient evidence" response and noting that some answers may be absent.

4 Results

The experiments reveal clear quality gains from pipeline enhancements (see results in Table 1), alongside recurring failure modes that concentrate in attribution, uncertainty handling, and workflow compliance. We unpack these insights in the following paragraphs.

Benchmark: effect sizes and interpretation. Increasing retrieval depth from top-k=3 to 5 produces a small but consistent uplift (R1: 3.63 → R2: 3.68), suggesting that many questions benefit from slightly broader evidence coverage.

Query rewriting is the strongest single lever (R2: 3.68 → R3: 3.96, +0.28). Qualitatively, rewriting helped most when the original question was underspecified, conversational, or used SDG shorthand: the rewriter expanded the query into a phrasing that better captures the user intent and leads to a more meaningful embedding.

Table 1: Run configurations and combined score (max 5). MM = multimodal.

Run	k	Rewrite	Rerank	MM	Lang	Combined
R1	3	N	N	N	EN	3.63
R2	5	N	N	N	EN	3.68
R3	5	Y	N	N	EN	3.96
R4	5	N	Y	N	EN	3.72
R5	5	Y	Y	N	EN	4.00
R6	5	Y	Y	N	FR	3.91
R7	5	Y	Y	Y	EN	4.04
R8	5	Y	Y	Y	FR	4.02

Reranking alone yields a smaller gain (R2: 3.68 → R4: 3.72), but combining rewriting and reranking is best among text-only variants (R5: 4.00). Qualitatively, without reranking, some retrieved chunks were only weakly relevant, and highly informative passages sometimes appeared deeper in the list (e.g., among ranks 6–10) than in the first few results. Reranking helps by reordering the candidate set so that the most on-topic evidence is surfaced early and passed to the generator.

Multimodal runs obtain the highest combined scores (R7: 4.04; R8: 4.02), driven mainly by higher relevance and better "referenceless" quality (answers remain coherent even when the judge does not see retrieved context), consistent with the generator being a larger, more capable model [21].

Switching internal prompt language (EN vs. FR) does not yield consistent improvements: for the strongest text-only configuration, FR is slightly worse, while multimodal differences are negligible. In our setting, this suggests that (English) rubric clarity may dominate over (French) lexical matching in intermediate prompting.

Benchmark: Source attribution remains the bottleneck. Across variants, provenance transparency is systematically lower than other dimensions. Source attribution is weakest on "easy" questions (e.g., definitional SDG questions), where models tend to blend retrieved snippets with background knowledge without explicitly marking which claims are supported by quotes. This matters for content analysis, because high apparent correctness does not imply that evidence can be traced or audited.

Held-out coding (50 videos): compliance, agreement, and where the system breaks. On held-out coding, the system typically follows the requested structured format and grounds most assigned codes in verbatim transcript excerpts. Hallucinated quotes were uncommon under the quote-first prompt.

The main compliance failures were (i) proceeding to full coding despite an explicit screening outcome indicating "partial"/"do not code" (i.e., violating workflow control), and (ii) occasional leakage from metadata (implicitly treating dataset SDG labels as truth even when the transcript suggests otherwise). A simple mitigation is to remove SDG labels from the model's input queries/prompts, forcing the system to infer SDG relevance from transcript evidence rather than from dataset annotation. More broadly, this highlights the

need for iterative evaluation and prompt refinement: small prompt changes can materially affect whether the model treats metadata as evidence.

Agreement with a small human-coded subset (10 videos) is strongest for low-ambiguity, single-label fields (explicit SDG mentions, broad stance/tone, simple format decisions). Divergences concentrate in nuance-heavy and variable-length fields: the model often under-reports stakeholders and responsibility attributions (multi-label under-selection), shows calibration drift on ordinal scales (e.g., agency/intensity), and overfits to the available taxonomy rather than selecting "unclear/insufficient evidence." In practice, these are precisely the dimensions where a human analyst's interpretive judgment is most valuable, supporting a hybrid "draft → review" workflow.

Claims and cross-video linking: extraction strong, comparison shallow. Claim/theme extraction is a consistent strength: it compresses noisy ASR captions into inspectable themes with supporting quotes, enabling rapid triage. In contrast, cross-video linking remains the main bottleneck. Links are often topical rather than mechanism- or framing-based; link types (e.g., "corroborate" vs. "extend mechanism") are sometimes overstated; and the model occasionally anchors on early retrieved candidates rather than the best match in the set. This suggests linking is bounded by both retrieval (candidate quality) and reasoning (typing and justification).

Thumbnail coding: vision is reliable, retrieval bounds usefulness. When a thumbnail is provided, the multimodal model reliably describes salient elements (text overlays, people, logos, nature imagery) and applies simple visual codes with few severe hallucinations. Errors are typically minor (small object additions or overly specific entity labels), indicating that most failures are in label mapping rather than perception.

However, independent thumbnail retrieval frequently surfaces generic "SDG poster" imagery (e.g., the SDG grid), which is on-theme but often uninformative for the analyst's specific query. As a result, the added value of multimodality is frequently bounded by image retrieval specificity and text–image alignment.

Uncertainty stress test: overfitting vs. fabrication, and prompt sensitivity. On plausible-but-unanswerable questions, the minimally constrained variant passed 3/5 cases; the two failures were (i) evidence stretching (overfitting to partially related snippets to force a specific answer) and (ii) evidence fabrication (inventing details not present in context).

With explicit guardrails, the system passed 5/5 cases, abstaining when the retrieved context did not support a specific answer and explicitly stating what evidence was missing. This indicates that epistemic safety is highly prompt-sensitive and that "insufficient evidence" behaviors are a practical guardrail for analyst-facing RAG.

5 Conclusion

Our findings suggest that RAG is a practical aid for SDG discourse auditing in long-form transcripts when outputs are anchored in verbatim evidence, but reliability hinges on explicit provenance guardrails. We synthesize the results into answers to RQ1–RQ3.

RQ1: Can RAG support evidence-grounded content analysis over long-form SDG transcripts? Overall, yes: across our tasks, the assistant is most reliable when the output is explicitly tied to verbatim transcript quotes (e.g., evidence-grounded lookup and claim/theme extraction) and when coding fields have clear lexical anchors. In this regime, RAG helps analysts triage long, noisy ASR captions into compact, checkable artifacts and navigate to relevant regions of long videos. However, the same experiments show that provenance transparency remains the core bottleneck: models often blend retrieved evidence with background knowledge, especially on seemingly "easy" questions, and cross-video linking tends to produce topical rather than mechanism- or framing-based comparators.

RQ2: How sensitive is the analytical system performance to RAG enhancement choices, and which improvements generalize across query types? Performance is sensitive to enhancement choice, with consistent, cross-task gains coming from enhancements. Across our query types, *query rewriting* is the strongest single lever: it improves retrieval relevance and downstream output quality at comparatively low cost. *LLM reranking* can provide additional gains when combined with rewriting by promoting the most on-topic chunks into the limited context budget, but it is more expensive and its benefit is limited. Multimodal variants obtain the best overall benchmark scores, most likely due to the larger size of the LLM.

RQ3: What are the dominant failure modes observed, and under which conditions do they occur? The dominant failure modes are (a) weak source attribution and evidence–knowledge boundary control, (b) overconfidence and insufficient uncertainty expression, (c) occasional metadata leakage, and (d) instruction non-compliance in structured workflows (e.g., proceeding to full coding despite a "partial"/"do not code" screening outcome). These issues cluster in predictable conditions: attribution and evidence–knowledge blending is most common on seemingly "easy" questions (e.g., definitional SDG prompts); workflow violations surface in multi-step coding pipelines; and metadata leakage appears when dataset annotations or labels are present in the model input. Encouragingly, our hallucination stress tests show that explicit guardrails can materially reduce fabrication.

Limitations. Our study has four limitations. First, the investigated corpus is predominantly French, so results may not directly transfer to other languages [5, 22]. Second, captions are ASR-generated and noisy [25], and the system has limited access to audio/visual signals, which can omit nuances that are important for coding [8]. Third, LLM-as-judge scores are subjective and can be brittle, so we interpret only aggregate trends [26]. Fourth, generalization is limited by the evaluation scope: held-out coding uses a 50-video sample and a mock codebook, so agreement patterns should be read as indicative [3].

Acknowledgments

This work was supported by the EU Horizon Europe program through the ELIAS project (No. 101120237). We thank Zhuosheng Huang for his help with data collection. We acknowledge the use of GPT to assist with the refinement of the text. The content, conclusions, and assertions in this paper are our own.

References

[1] Yosra Abdessamed, Shadi Rezapour, and Steven Wilson. 2024. Identifying Narrative Content in Podcast Transcripts. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, St. Julian's, Malta, 2631–2643. doi:10.18653/v1/2024.eacl-long.161

[2] Subhash Abhayawansa. 2011. A methodology for investigating intellectual capital information in analyst reports. *Journal of Intellectual Capital* 12, 3 (2011), 446–476. doi:10.1108/14691931111154733

[3] Maria Becker, Mirko Sommer, Lars Tapken, Yi Wen Teh, and Bruno Brocai. 2025. The Moralization Corpus: Frame-Based Annotation and Analysis of Moralizing Speech Acts across Diverse Text Genres. arXiv:2512.15248 [cs.CL] doi:10.48550/arXiv.2512.15248

[4] Frank Biermann, Thomas Hickmann, Carole-Anne Sénit, Marianne Beisheim, Steven Bernstein, Pamela Chasek, Leonie Grob, Rakhyun E. Kim, Louis J. Kotzé, Måns Nilsson, Andrea Ordóñez Llanos, Chukwumerije Okereke, Prajal Pradhan, Rob Raven, Yixian Sun, Marjanneke J. Vijge, Detlef van Vuuren, and Birka Wicke. 2022. Scientific evidence on the political impact of the Sustainable Development Goals. *Nature Sustainability* 5, 9 (Sept. 2022), 795–800. doi:10.1038/s41893-022-00909-5

[5] Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, Sha Li, Manling Li, Michael JQ Zhang, Eunsol Choi, Mor Geva, Peter Hase, and Heng Ji (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 177–188. doi:10.18653/v1/2024.knowllm-1.15

[6] Chroma. n.d.. Chroma Clients: PersistentClient. https://cookbook.chromadb.dev/core/clients/

[7] Teyl Engstrom, Jenny Strong, Clair Sullivan, and Jason D. Pole. 2022. A Comparison of Leximancer Semi-automated Content Analysis to Manual Content Analysis: A Healthcare Exemplar Using Emotive Transcripts of COVID-19 Hospital Staff Interactive Webcasts. *International Journal of Qualitative Methods* 21 (2022), 16094069221118993. doi:10.1177/16094069221118993

[8] Sahar Fazeli, Judith Sabetti, and Manuela Ferrari. 2023. Performing Qualitative Content Analysis of Video Data in Social Sciences and Medicine: The Visual-Verbal Video Analysis Method. *International Journal of Qualitative Methods* 22 (2023). doi:10.1177/16094069231185452

[9] Chenhan Fu, Guoming Wang, Juncheng Li, Wenqiao Zhang, Rongxing Lu, and Siliang Tang. 2025. ITERATE: Image-Text Enhancement, Retrieval, and Alignment for Transmodal Evolution with LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 1365–1376. https://aclanthology.org/2025.coling-main.91/

[10] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The Llama 3 Herd of Models. (2024). arXiv:2407.21783 [cs.AI] doi:10.48550/arXiv.2407.21783

[11] Iñaki Heras-Saizarbitoria, Laida Urbieta, and Olivier Boiral. 2022. Organizations' engagement with sustainable development goals: From cherry-picking to SDG-washing? *Corporate Social Responsibility and Environmental Management* 29, 2 (March 2022), 316–328. doi:10.1002/csr.2202

[12] Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. 2025. VideoRAG: Retrieval-Augmented Generation over Video Corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 21278–21298. doi:10.18653/v1/2025.findings-acl.1096

[13] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth J. F. Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2021. TREC 2020 Podcasts Track Overview. (March 2021). arXiv:2103.15953 [cs.IR] doi:10.48550/arXiv.2103.15953

[14] Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. 2024. jina-clip-v2: Multilingual Multimodal Embeddings for Text and Images. (Dec. 2024). arXiv:2412.08802 [cs.CL] doi:10.48550/arXiv.2412.08802

[15] Marvin Limpijankit and John Kender. 2025. Detecting Cultural Differences in News Video Thumbnails via Computational Aesthetics. arXiv:2505.21912 [cs.CY] doi:10.48550/arXiv.2505.21912

[16] Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. 2024. Video-RAG: Visually-aligned Retrieval-Augmented Long Video Comprehension. *arXiv preprint arXiv:2411.13093* (2024). doi:10.48550/arXiv.2411.13093

[17] Francesca Manes-Rossi and Giuseppe Nicolò. 2022. Exploring sustainable development goals reporting practices: From symbolic to substantive approaches—Evidence from the energy sector. *Corporate Social Responsibility and Environmental Management* 29, 5 (Sept. 2022), 1799–1815. doi:10.1002/csr.2328

[18] Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. 2025. A Survey of Multimodal Retrieval-Augmented Generation. (2025). arXiv:2504.08748 [cs.IR] doi:10.48550/arXiv.2504.08748

[19] Mistral AI. 2025. Mistral-Small-3.1-24B-Instruct-2503 (model card). https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503

[20] Ana I. Moreno and John M. Swales. 2018. Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes* 50 (April 2018), 40–63. doi:10.1016/j.esp.2017.11.006

[21] Jingjie Ning, Yibo Kong, Yunfan Long, and Jamie Callan. 2025. Less LLM, More Documents: Searching for Improved RAG. arXiv:2510.02657 [cs.IR] doi:10.48550/arXiv.2510.02657

[22] Jeonghyun Park and Hwanhee Lee. 2025. Investigating Language Preference of Multilingual RAG Systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 5647–5675. doi:10.18653/v1/2025.findings-acl.295

[23] Alessia Patuelli and Fabio Saracco. 2023. Sustainable development goals as unifying narratives in large UK firms' Twitter discussions. *Scientific Reports* 13, 1 (April 2023), 7017. doi:10.1038/s41598-023-34024-y

[24] Anna Scaini, Chiara Scaini, Jay Frentress, Georgia Destouni, and Stefano Manzoni. 2021. Linking the 2030 Agenda for Sustainable Development to Research, Newspapers, and Governance: The Case of the Last Free-Flowing Alpine River. *Frontiers in Environmental Science* 9 (2021), 553822. doi:10.3389/fenvs.2021.553822

[25] Rosy Southwell, Samuel Pugh, E. Margaret Perkoff, Charis Clevenger, Jeffrey Bush, Rachel Lieber, Wayne Ward, Peter Foltz, and Sidney D'Mello. 2022. Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms. In *Proceedings of the 15th International Conference on Educational Data Mining*, Antonija Mitrovic and Nigel Bosch (Eds.). International Educational Data Mining Society, Durham, United Kingdom, 302–315. doi:10.5281/zenodo.6853109

[26] Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large Language Models are Inconsistent and Biased Evaluators. arXiv:2405.01724 doi:10.48550/arXiv.2405.01724

[27] Wuyou Sui, Anna Sui, and Ryan E. Rhodes. 2022. What to watch: Practical considerations and strategies for using YouTube for research. *Digital Health* 8 (Sept. 2022), 20552076221123707. doi:10.1177/20552076221123707

[28] United Nations General Assembly. 2015. Transforming our world: the 2030 Agenda for Sustainable Development. Resolution A/RES/70/1. https://docs.un.org/en/A/res/70/1

[29] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. (Feb. 2024). arXiv:2402.05672 [cs.CL] doi:10.48550/arXiv.2402.05672

[30] Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2024. Searching for Best Practices in Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 17716–17736. doi:10.18653/v1/2024.emnlp-main.981

[31] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-Augmented Generation for AI-Generated Content: A Survey. (2024). arXiv:2402.19473 [cs.CV] doi:10.48550/arXiv.2402.19473