Seeing Through the MIRAGE: Evaluating Multimodal Retrieval Augmented Generation

Anonymous ACL submission

Abstract

We introduce MIRAGE, an evaluation framework for retrieval-augmented generation (RAG) from multimodal sources. As audiovisual media becomes a prevalent source of information online, it is essential for RAG systems to integrate information from these sources into generation. However, existing evaluations for RAG are text-centric, limiting their applicability to multimodal settings. MIRAGE is a claim-centric approach to multimodal RAG evaluation, consisting of INFOF1, which assesses factuality and information coverage, and CITEF1, which assesses citation support and completeness. We show that, when applied by humans, MIRAGE strongly aligns with extrinsic judgments of output quality. We additionally introduce an automatic implementation of MIRAGE as well as multimodal variants of three prominent text-based RAG metrics—ALCE, ARGUE, and RAGAS—demonstrating the limitations of text-centric work and laying the groundwork for automatic evaluation. We release open-source implementations[1] and outline evaluation methods for multimodal RAG.

1 Introduction

Online information is increasingly consumed through audiovisual media—from firsthand video footage of natural disasters to professional news coverage of major political events. It is thus essential that systems for retrieval-augmented generation (RAG) be able to integrate information from *any* modality and provide accurate citations to multimodal sources. However, work on long-form text generation from multimodal sources (Krishna et al., 2017; Liu et al., 2023; Lin et al., 2024; Martin et al., 2025) relies on text-only metrics for evaluation (e.g., Lin, 2004; Zhang et al., 2019), which do not verify information against non-text sources.

Subclaim-level factuality metrics like FActScore (Min et al., 2023) verify outputs by decomposing them into subclaims and verifying each subclaim against a knowledge source. However, FActScore and its variants (Song et al., 2024) are designed for single-source, text-to-text verification, and do not handle citations, multiple sources, or multimodal evidence. RAG evaluation requires more: systems must not only produce factual text, but attribute information to the correct sources, which may be images, audio, or video, in addition to text.

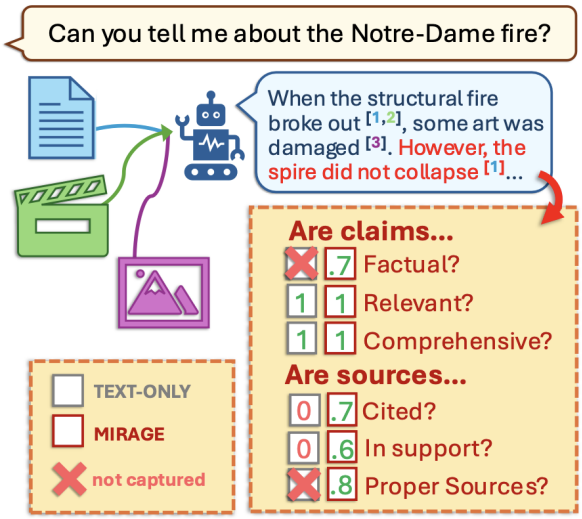

Figure 1: MIRAGE asseses predictions at the subclaim level, evaluating information factuality and coverage, as well as citation support and attribution, enabling RAG evaluation grounded in text, image, audio, and video.

Existing text-only RAG metrics—ALCE (Gao et al., 2023), ARGUE (Mayfield et al., 2024), and RAGAS (Es et al., 2024)—address citation quality and information coverage, but are designed for text passages and operate at the sentence level, precluding fine-grained (i.e., subclaim-level) factuality judgments. Extending these metrics to multimodal sources is not straightforward: multimodal inference introduces a much heavier computational burden (e.g., fitting multiple videos in context), novel challenges relating to evidential support (e.g., how to comprehensively extract facts from videos), and

[1] Data and code upon publication.

(sub)claim grounding that requires cross-modal reasoning over visual and auditory signals. We show that naively adapting text-only RAG metrics to video yields poor correlation with human judgments (§6), confirming the need for a framework designed specifically for multimodal RAG.

We introduce MIRAGE, a framework for **M**ult**i**modal **R**etrieval **A**ugmented **G**eneration **E**valuation that extends subclaim-level factuality assessment to the multimodal, multi-source, citation-attributed setting. Core to MIRAGE is the notion that all information—whether represented in text, image, audio, or video—can be decomposed into subclaims, grounding all modalities in a common unit of evaluation. Our framework consists of INFOF1, which measures the factuality (precision) and coverage (recall) of information in generated text, and CITEF1, which measures whether cited sources support their associated subclaims (precision) and whether recalled information is properly attributed (recall). Both metrics can be evaluated in reference-based or reference-free settings, enabling MIRAGE to be used across an array of different tasks (e.g., summarization, VQA).

We evaluate on two multimodal RAG datasets, WIKIVIDEO (Martin et al., 2025) and MAGMaR, focusing on audiovisual (video) sources. Comparing against three types of human judgments, we find that MIRAGE has the highest agreement with overall quality judgments, and that subclaim-level precision (INFOP, CITEP) is the only automatic metric that consistently correlates with grounding judgments across both datasets, while other metrics have low or negative correlation.

In summary: (1) We propose MIRAGE, extending subclaim-level factuality assessment to multimodal, multi-source, citation-attributed RAG; (2) we present multimodal versions of three prominent text-only RAG metrics, revealing their limitations; and (3) we evaluate on two datasets with three types of human judgments, showing that MIRAGE generalizes where natively text-based metrics do not.

2 Related Work

Retrieval-Augmented Generation has been widely studied in text-only settings (TextRAG; Lewis et al., 2020; Barham et al., 2023; Gupta et al., 2024; Han et al., 2024; Lawrie et al., 2024). Recent work has extended RAG to other modalities, including documents with visual elements (Cho et al., 2024), speech (Min et al., 2025; Chen et al., 2025), and video (VideoRAG; Jeong et al., 2025; Martin et al., 2025; Ren et al., 2025). We focus on the VideoRAG task as proposed in Martin et al. (2025), as it best mirrors the common TextRAG setting: long-form generation from multiple videos with citations. Video subsumes both visual and audio modalities, making it a challenging testbed. We also validate that our evaluation generalizes independently to text, audio, and video.

Claim Decomposition is a core component of factuality assessment. An important precursor is the Pyramid method (Nenkova and Passonneau, 2004), which abstracts summaries into Summary Content Units (SCUs)—minimal, semantically atomic propositions used as the basis for evaluation. SCUs introduced the principle that content should be assessed at the level of individual facts rather than full sentences, anticipating the subclaim decomposition central to modern factuality metrics. Building on this principle, FActScore (Min et al., 2023) and VeriScore (Song et al., 2024) decompose predictions into subclaims and verify each one against a single text source, arguing that subclaims enable more straightforward factuality assessment than full sentences in virtue of their atomic propositional content. Subsequent work has attempted to refine the appropriate conception of *atomicity* for subclaims (Gunjal and Durrett, 2024; Wanner et al., 2024a,b) and has extended the core methodology to images (Jing et al., 2024). However, these metrics evaluate a prediction against a single source and do not assess citations, which requires verification across multiple sources. MIRAGE builds on this foundation but addresses a different setting: multi-source generation with citation attribution, where each subclaim must be factual and cited by the correct source.

3 TextRAG Metrics

Before introducing MIRAGE, we first revisit how RAG is assessed in text-only settings (TextRAG). We look at three prominent metrics—ALCE (Gao et al., 2023), ARGUE (Mayfield et al., 2024), and RAGAS (Es et al., 2024)—providing a foundation for RAG evaluation as well as the modifications to these metrics for handling multimodal sources. Implementation details for each are in Appendix D.

Preliminaries We define a *document* as any single unit of evidence, regardless of modality (e.g., a text passage, image, audio clip, or video). We

define a *topic* as the subject matter of one or more *documents*. The documents used in generation are the *context*. A *prediction* is text produced by a generation system, decomposed into *subclaims*—atomic declarative statements that may or may not be true. A *reference* is a collection of preexisting knowledge on a query topic, such as a knowledge base (e.g., Wikipedia), a set of nuggets (Voorhees, 2004), or a (human-written) reference text (Martin et al., 2025). We prefix *predicted* or *reference* to denote the origin of derived objects (e.g., *predicted subclaim*, *reference subclaim*). A scoring function $\mathsf{s}(p, h)$ measures support between a (premise, hypothesis) pair via entailment or LLM judgment.

3.1 ALCE

ALCE (Gao et al., 2023) aims to capture three dimensions of predicted text quality: information correctness, citation quality, and fluency.[2]

Correctness ALCE introduces both recall and precision metrics to evaluate Correctness. Recall is a natural language inference- (NLI-)based judgment between reference subclaims and the predicted output; precision is exact match between predicted and reference answers. We do not implement precision, as exact match is unsuitable for long-form generation. We do not attempt to relax exact match in our implementation, as this would reduce to a metric like BERTScore, which we already include in our evaluation. Claim Recall requires no modification for multimodality, as verifying reference subclaims against a predicted text is independent of the modality of the sources.

Citation Quality ALCE also uses recall and precision metrics to evaluate Citation Quality. Citation Quality Recall assess whether all the sources cited by a sentence, jointly considered (i.e., concatenated), support the citing sentence. Citation Quality Precision penalizes irrelevant citations by assessing whether a sentence remains (jointly) supported by its cited sources when one of them is removed. Both precision and recall involve comparing a sentence against the concatenation of multiple text passages. In this work, we instead consider multiple multimodal documents.

Citation Quality has two limitations: (1) sentences typically express multiple subclaims, making support harder to evaluate than with subclaims, and (2) concatenating multiple documents is computationally infeasible for long multimodal sources and unsupported by current training data.

3.2 ARGUE

Similar to ALCE, ARGUE (Mayfield et al., 2024) assesses generations for information coverage and for citation support. For the former, ARGUE adopts a *nugget*-based evaluation, assessing what proportion of a set of question-answer pairs (*nuggets*) about a target topic are attested by the generated text (Nugget Coverage). For the latter, ARGUE calculates the support between individual (sentence, citation) pairs (Sentence Support). We rely on the AUTO-ARGUE implementation of ARGUE from Walden et al. (2025a).

Nugget Coverage is the proportion of nugget questions correctly answered by the predicted text, aggregated over sentences. This is similar to ALCE's Claim Recall, but focused on recall of QA pairs rather than of subclaims. Since nugget coverage requires only these nuggets and the generated text for assessment, the only thing needed for our evaluation is to create the collection of nuggets.

Sentence Support is the proportion of predicted sentences attested individually by each of their citations. In focusing on sentences rather than subclaims, Sentence Support runs into the same challenge as ALCE's Citation Quality evaluation in having to simultaneously assess the support of multiple subclaims expressed by a single sentence.

3.3 RAGAS

RAGAS (Es et al., 2024) evaluates three dimensions of predicted text quality: faithfulness to the retrieved context (Faithfulness), relevance of the text to an information need (Answer Relevance), and relevance of the retrieved context to the information need (Context Relevance).

Faithfulness For Faithfulness, sentences from the predicted text are decomposed into "statements" (i.e., subclaims[3]) and are compared against all documents in the retrieved context to obtain a binary judgment of support. In our implementation, we apply the same claim decomposition method as we use in MIRAGE to obtain these statements. We evaluate statement support against the documents

[2] We do not explore fluency in this work, as it is not a common failure mode of modern LLMs.

[3] Es et al. (2024) create one or more statements from a sentence, which we interpret as subclaim decomposition.

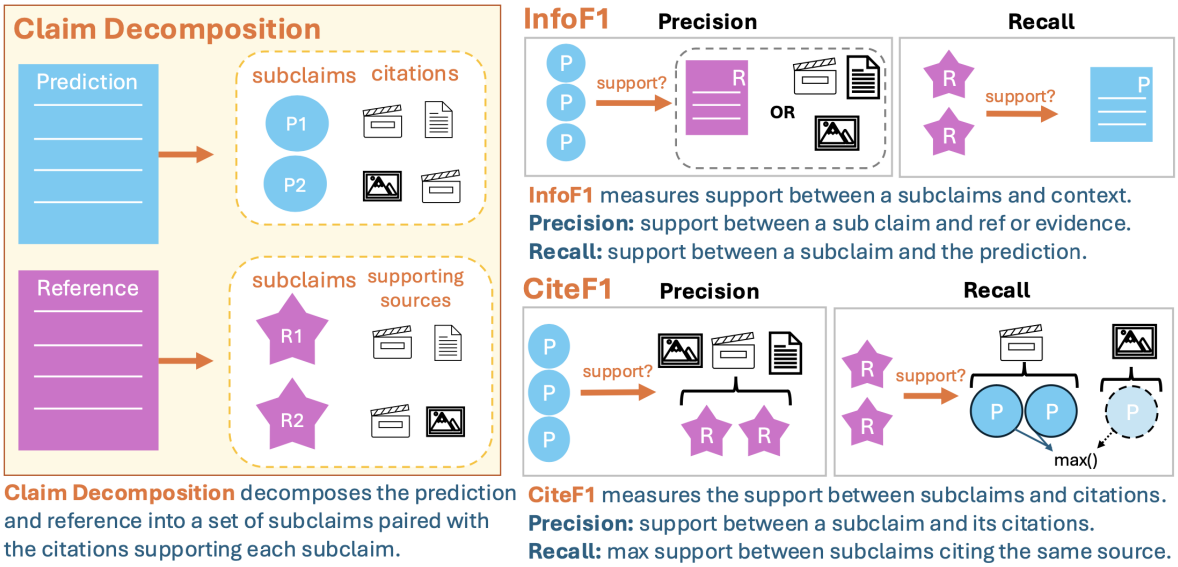

Figure 2: MIRAGE decomposes generated text into subclaims to assess two dimensions of quality: INFOF1, which measures factuality and information coverage, and CITEF1, which measures citation precision and completeness. The framework enables consistent, claim-level evaluation across text, image, audio, and video modalities.

in context one-by-one, rather than jointly, owing to context limitations.

Answer Relevance For Answer Relevance, RAGAS first uses an LLM to predict a query that could have yielded the predicted text as a response. Answer Relevance is then computed as the cosine similarity between the embeddings of the predicted and actual queries, given some embedding model.

Context Relevance RAGAS uses an LLM to extract sentences from the retrieved context that are judged to be essential to answering the query, computing the Context Relevance as $\frac{\text{extracted sentences}}{\text{total context sentences}}$. To adapt this for multimodal documents (videos, in our experiments), we extract query-relevant observations from each document and (separately) elicit a detailed VLM summary of the same document to approximate the denominator.

3.4 Summary

The three metrics above share core limitations that hinder their applicability in multimodal domains: ALCE and ARGUE assess information at the sentence level, making fact verification less precise. Additionally, ALCE relies on concatenating multiple sources, which is infeasible with multiple long multimodal documents, and RAGAS assumes an exhaustive extraction of observations from sources—an impracticable expectation in the multimodal domain. MIRAGE addresses these issues by decomposing predictions into subclaims across all evaluation dimensions, enabling more granular assessments of factuality and citation quality. Its two components, INFOF1 and CITEF1, can be computed in both reference-based and reference-free settings, providing a scalable, claim-centric framework for evaluating multimodal RAG.

4 MIRAGE

MIRAGE (Figure 2) is a subclaim-centric evaluation framework that evaluates two dimensions of predictions: the presented **information** and the use of **citations**. INFOF1 measures both the proportion of predicted subclaims that are factual (precision) as well as the proportion of a set of reference subclaims that are supported by the predicted text (recall). CITEF1 measures whether subclaims are supported by their cited sources (precision) and whether reference subclaims attested by the prediction are attributed to a source that supports them (recall). Appendix B provides further variants of MIRAGE for single-source inference and subclaim and source importance weighting.

Subclaim Decomposition A core step in MIRAGE is decomposing sentences into subclaims. We decompose predictions one sentence at a time using an LLM prompted to create a set of *decontextualized* subclaims (Gunjal and Durrett, 2024; Wanner et al., 2024a), ensuring that subclaims are atomic but independently verifiable. In our experiments, we use manually-authored reference subclaims, though automatic decomposition can be applied to other reference forms as well.

Reference Evaluations A key aspect of MIRAGE is the ability to evaluate in both reference-free (Collection) and reference-based (Reference) settings. In the Collection setting, subclaims are verified directly against the documents themselves. In the Reference setting, each document is mapped to a text-based proxy for its information content (e.g., a human-written article or set of manually authored subclaims). This makes Reference evaluation more computationally efficient, as it avoids multimodal inference, but introduces the limitation that human-authored references are inherently non-exhaustive: factual subclaims omitted from the reference will be penalized.

4.1 INFOF1: Evaluating Information Quality

We evaluate information quality using an F1 score over two metrics: INFOP, which assesses subclaim factuality, and INFOR, which assesses information coverage. We refer to the resulting metric as INFOF1.

4.1.1 Information Factuality (INFOP)

We evaluate the factuality of a predicted text P against a reference set of evidence documents $R = \{r_1, \ldots, r_n\}$. We sentence-split the predicted text and then decompose each predicted sentence into a set of subclaims (C_P). These subclaims are then scored for support against R using a scoring function $\mathtt{s}$, where INFOP is defined as the proportion of predicted subclaims that are supported, weighted by the support strength given by $\mathtt{s}$:

$$\text{precision}(C_P, R) = \frac{1}{|C_P|} \sum_{c \in C_P} \mathtt{s}(c, R) \quad (1)$$

We propose two variations of INFOP, Collection Precision and Reference Precision, which differ in how they construct R.

Collection Precision defines R as the set of relevant documents for the query topic. It is thus similar to RAGAS Faithfulness, which evaluates support against the retrieved documents used as generation context, although our experiments rely on the gold set of relevant documents for each topic. In principle, R could be a very large collection (e.g., a knowledge base), which would make it akin to a multimodal FActScore (Min et al., 2023).

Reference Precision defines R as the reference text or reference subclaims for the topic, evaluating support between predicted subclaims and the text proxy rather than the documents directly.

4.1.2 Information Coverage (INFOR)

Information Coverage (INFOR) is the proportion of subclaims in R that are attested in P. It is exactly analogous to INFOP, but with the roles of R and P reversed. If R is given as a complete reference text, we first sentence-split R and then decompose each sentence into a set of subclaims C_R.[4]

$$\text{recall}(C_R, P) = \frac{1}{|C_R|} \sum_{c \in C_R} \mathtt{s}(c, P) \quad (2)$$

INFOR is the same as ALCE Claim Recall, with both checking if claims in some reference are attested by the prediction.

A *collection* vs. *reference* distinction, analogous to the one for INFOP, can be made here for INFOR. In our experiments, we use the reference version of INFOR, as the collection version imposes impractical computational costs and task complexity.

4.2 CITEF1: Evaluating Citations

CITEF1 measures how effectively P grounds information in cited sources. Like INFOF1 (and unlike ALCE and ARGUE), CITEF1 works with subclaims. It is an F1 score over CITEP—the proportion of subclaims supported by their cited sources—and CITER—the proportion of subclaims that properly attribute their sources.

4.2.1 Citation Support (CITEP)

CITEP is the proportion of predicted subclaims that are supported by their cited documents, weighted by the strength of that support. We first decompose P into a set of subclaims C_P and create a mapping from each subclaim c_i^P to the set of documents D_i that are cited by the sentence from which c_i^P was decomposed; i.e., $\{(c_i, D_i) \mid c_i^P \in C_P\}$. Given the support scoring function $\mathtt{s}$, we compute:

$$\text{precision}(C_P, D) = \frac{1}{|C_P|} \sum_{c_i^P \in C_P} \mathtt{s}(c_i^P, D_i) \quad (3)$$

As with INFOP, CITEP has *collection* and *reference* variants, which differ in the form of the cited documents used in evaluation.

Collection Precision evaluates support between subclaims in P and their associated cited documents directly—i.e., the raw text (or video) content. This is similar to ARGUE's Sentence Support; however, instead of evaluating support between a sentence and citation, we evaluate support between a *subclaim* and a cited document.

[4] If R is given as a set of subclaims, we use these directly.

Reference Precision evaluates support between subclaims in P and subclaims obtained from the associated cited documents, rather than the raw document content. In our experiments, we create a mapping from each cited document d_i to the set of reference subclaims C_{R,d_i} attested by d_i: $\{(d_i, C_{R,d_i}) \mid P \text{ cites } d_i\}$, and evaluate support against C_{R,d_i} rather than the document itself:

$$\mathsf{s}(c_i^P, C_{R,d_i}) \tag{4}$$

Prior metrics, ARGUE and ALCE, evaluate full sentences against their citations, but this requires the sentence to be fully supported by *each* document (ARGUE) or the cited documents (ALCE). This overpenalizes multi-premise sentences for containing un- or previously supported subclaims alongside supported ones, generally rewarding unnatural texts with short, single-subclaim sentences.

4.2.2 Proper Attribution (CITER)

CITER is the proportion of *reference* subclaims (C_R) that are attested by P via sentences that cite valid sources—i.e., sources that *also* attest those subclaims. We first decompose the reference R into subclaims and create a mapping from each subclaim c_i^R to the set of documents D_i in the collection that (individually) attest that subclaim: $\{(c_i^R, D_i) \mid c_i^R \in R\}$. Next, we create another mapping from each document d_j that is cited in P to the set S_j of sentences in P that cite d_j: $\{(d_j, S_j) \mid d_j \in P\}$. We then compose these two mappings to obtain the proportion of reference subclaims that are both attested in P *and* supported by a valid source citation:

$$\text{recall}(C_R, P, D) = \frac{1}{|C_R|} \sum_{c_i^R \in C_R} \max_{d_j \in D_i} \left(\max_{s_k \in S_j} \mathsf{s}(c_i^R, s_k) \right) \tag{5}$$

We take the max support score over all documents $d_j \in D_i$ that attest the target subclaim (c_i^R)—and over all sentences s_k in P that cite d_j—because we care about the *best* support that P provides for each c_i^R *anywhere in the text*.

Unlike INFOR, CITER does not have the distinction between collection and reference settings, as CITER relies on the set of *reference* subclaims, ensuring that they are recalled with a valid source.

5 Grounding Claims in Multimodal Data

An important component of MIRAGE is the ability to ground claims in multimodal data. We evaluate claim grounding across four modality settings—Text (UNLI (Chen et al., 2020); PeopleProfile (Walden et al., 2025b)), Audio (CLOTHO (Drossos et al., 2020; Deshmukh et al., 2024); WIKIVIDEO-Audio), Video, and Audiovisual (WIKIVIDEO)—using audio-language (Goel et al., 2025; Chu et al., 2024), vision-language (Bai et al., 2025; Team, 2026), and omnimodal models (Xu et al., 2025; Zhang et al., 2026). Full details are in Appendix A.

Model	P	Text	Audio	Video	AudVid
AF	7B	50.9	60.0	–	–
Q2-A	7B	50.9	77.9	–	–
Q2.5-V	7B	64.0	–	22.1	–
Q3.5	9B	60.1	–	76.3	–
Q2.5-O	7B	65.3	81.9	50.5	45.2
CLUE	3B	63.9	89.8	74.6	70.1

Table 1: Accuracy of claim grounding methods. P: Parameters. AF: Audio Flamingo. Q: Qwen, A-Audio VL-Vision Language, O-Omni.

In Table 1, we find that most methods struggle with claim verification: modality-specific models perform decently in their respective modalities but lack the coverage needed for general-purpose multimodal evaluation, while omnimodal models offer broader coverage at varying quality. CLUE achieves comparable performance in each modality, making it suitable as a general-purpose omnimodal verifier. We report additional human alignment with other models in Appendix C.

6 Alignment With Human Judgments

Evaluation Setup We evaluate MIRAGE on two datasets: WIKIVIDEO (Martin et al., 2025) and the MAGMaR test set. For MAGMaR, we use the predictions from: Bhosale et al. (2026), Chakraborty et al. (2026), and Yan et al. (2026). For WIKIVIDEO, we evaluate predictions with three CAG system variations (no videos, oracle videos, retrieved videos) on 10 topics and collect our own set of human judgments.

Collection of Human Judgments We collect three human judgments.[5] (1) **Extrinsic Quality Judgments (EQJs)**. Human annotators are given information about the topic and are asked to give a 1-5 Likert score considering the following attributes (in order of importance): factuality, adequacy, coherence, relevancy, and fluency. (2) **In-**

[5] See Appendix E for instructions.

	EQJ 1	EQJ 2	EQJ 3
ICJ 1	49.0	61.3	54.5
ICJ 2	60.5	81.2	72.7
ICJ 3	65.4	76.3	70.8
GJ	18.2	-5.5	23.0

Table 2: Kendall's Tau agreement between Human-annotated MIRAGE (ICJ[1-3]/GJ) and human annotated quality judgments (EQJ[1-3]).

trinsic Claim Judgments (ICJs). Human annotators are given the prediction, predicted subclaims, reference, and reference subclaims, and are asked to annotate whether or not a claim is supported by the other text. This directly mirrors INFOF1-Ref: the scores are what INFOF1 would produce if a human performed the verification instead of a model. (3) **Grounding Judgments (GJs)**. Human annotators are given all the relevant videos for a topic and the predicted subclaims and asked to annotate whether or not the subclaims are supported by (grounded in) the videos. This directly mirrors INFOP and CITEP: the scores are what these metrics would produce with human verification.

In the following paragraphs, we will summarize the agreements and findings between human judgments and automatic methods, but leave discussion on *how* to evaluate multimodal RAG to section 7. The agreement is a Kendall's Tau (Kendall, 1938) between the human and metric ranking of systems.

6.1 Human-Annotated MIRAGE

ICJs and GJs are human-computed instances of MIRAGE: they apply the same claim-level precision and recall, but with human annotators performing the verification step rather than a model. In Table 2, we evaluate whether these human-computed MIRAGE scores agree with the extrinsic quality judgments (EQJs), which reflect holistic human preferences independent of any specific metric.

We find that INFOF1*, when computed by humans, aligns strongly with EQJs*. This demonstrates that with human-level verification performance, MIRAGE captures human quality preferences, validating the framework's design independently of any particular automatic verifier. We also observe that grounding judgments (GJs) show inconsistent agreement with EQJs, aligning with some annotators (EQJ 1 and 3) but not others. This suggests that groundedness in the source material contributes to overall quality but is not fully captured by quality judgments alone. This result reinforces that *metrics targeting quality or information coverage may not capture groundedness unless they explicitly verify against the sources of information*.

6.2 Automatic Evaluations

We next evaluate the automatic implementations of MIRAGE and TextRAG metrics, reporting agreement with the human judgments on WIKIVIDEO (Table 3, Table 4) and MAGMaR (Table 5).

MIRAGE best captures human preference. MIRAGE achieves the strongest and most consistent agreement with human judgments across both datasets and judgment types. On WIKIVIDEO, INFOF1 (Table 3) achieves strong alignment with both EQJs and ICJs, while CITEF1 (Table 4) similarly aligns well with EQJs. Most notably, MIRAGE is the only metric that achieves positive agreement with GJs—all other metrics produce near-zero or negative correlations, indicating they fail to capture whether predictions are actually supported by the source material. This pattern generalizes to MAGMaR (Table 5), where MIRAGE maintains positive agreement with quality judgments across both information and citation dimensions, confirming that its design transfers across tasks and annotation setups.

Subclaims outperform sentence-level. Comparing MIRAGE's claim-level evaluation against the sentence-level approaches of ALCE, ARGUE, and RAGAS reveals a consistent pattern: evaluating at the sentence level reduces agreement with human judgments. Sentence-level metrics like ALCE Citation Quality and ARGUE Sentence Support show negative correlation with GJs and weaker correlations with EQJs than their claim-level MIRAGE counterparts (Table 4). This result further demonstrates atomic subclaims as the appropriate granularity for evaluation against multimodal sources over sentences, adding to the findings of claim literature in text (Min et al., 2023).

Surface-level metrics fail to generalize and ground. On WIKIVIDEO, where high-quality human-written articles exist for the topics, ROUGE and BERTScore achieve reasonable agreement with quality (EQJ) and information (ICJ). However, they fail to capture grounding, producing near-zero or negative correlation with GJs, missing a fundamental requirement of RAG evaluation: verifying information against its sources. On MAGMaR,

Eval	Ano	INFOP		INFOR	ALCE	ARGUE	RAGAS		R-L	BS
		Ref.	Col.	Ref.	Clm. Rec.	Nugg. Cov.	Faith	Ans. Rel.		
EQJ	1	19.7	32.8	72.7	6.7	-1.8	15.2	4.8	55.8	52.3
	2	18.2	31.0	31.2	13.3	10.0	23.2	23.3	71.2	81.2
	3	17.8	28.2	13.0	14.8	-8.5	16.3	-1.5	41.2	74.2
ICJ	1	30.0	18.5	54.5	34.8	0.0	11.8	44.8	66.7	70.0
	2	40.0	36.7	64.5	31.5	13.3	20.0	33.3	70.0	86.7
	3	40.0	8.5	66.0	18.5	20.0	26.7	23.3	68.2	78.2
GJ	-	6.7	10.0	1.5	-5.2	-13.3	3.0	-26.7	-11.5	0.0

Table 3: Kendall's Tau with all human judgments against the metrics that evaluate information. R-L: ROUGE-L, BS: BERTScore, EQJ: Extrinsic Quality Judgment, ICJ: Intrinsic Claim Judgment, GJ: Grounding Judgment

Eval	CITEP		CITER	AC	ARG	RG
	R	C	R	CQ	SS	CR
EQJ	11.5	-1.8	38.2	-20.0	-15.0	32.7
	33.0	15.2	51.2	10.0	1.8	23.2
	14.8	16.3	39.7	8.2	-1.8	13.2
GJ	0.0	26.7	1.5	-8.5	-3.3	3.0

Table 4: Kendall's Tau for all human judgments against citation metrics. AC: ALCE, ARG: ARGUE, RG: RAGAS. R: Reference, C: Collection.

Eval	INFOP		INFOR	CITEP		CITER	R-L	BS
	R	C	R	R	C	R		
EQJ	41	4	31	26	26	28	-4	9
	65	-12	38	50	3	48	-1	7
	52	14	25	38	22	18	16	15

Table 5: Kendall's Tau on MAGMaR between MIRAGE, Rouge (R-L), and BERTScore (BS), and human judgments. R: Reference, C: Collection.

these metrics produce low agreement with EQJs, with ROUGE producing slight negative agreement. These results confirm that surface-level metrics are brittle proxies for quality, dependent on high-quality references and unable to assess grounding.

Automatic verification remains a bottleneck. While MIRAGE's framework design is validated by the strong agreement between human-computed scores and quality judgments (Table 2), the automatic variants do not capture the same level of agreement with extrinsic quality. As shown in section 5, the best omnimodal verifier achieves only modest accuracy on audiovisual verification. The gap between human-annotated MIRAGE and its automatic counterpart emphasizes (1) that MIRAGE with human-level grounding can achieve strong agreement and (2) the need for future work in improving the verification of claims in multimodal evidence.

7 How to Evaluate Multimodal RAG

We offer two recommendations for evaluating multimodal RAG. For quick diagnostics or compute-constrained settings, we recommend reference-based INFOF1 and CITEF1 as fast checks for factuality and grounding. For comprehensive evaluation, we recommend collection-based INFOF1 and CITEF1, which avoid the limitations of surface-level and sentence-level metrics shown above. ROUGE and BERTScore remain reasonable proxies when high-quality references exist, and RAGAS Faithfulness and ALCE Claim Recall closely mirror components of MIRAGE. Evaluating answer relevance, as RAGAS Answer Relevance does (although not perfectly), also remains important. **For MIRAGE, we recommend reporting Precision and Recall alongside F1, as they are fundamental; the F1 is derived.**

8 Conclusion

We introduce MIRAGE, a subclaim-centric framework for evaluating multimodal RAG. Through INFOF1 and CITEF1, MIRAGE provides consistent, claim-level assessment of factuality, information coverage, and citation quality across text, image, audio, and video. We show that MIRAGE aligns with human quality judgments where existing text-centric and sentence-level metrics do not. As future work continues to improve omnimodal verification, MIRAGE will enable scalable and reliable evaluation for multimodal generation.

Limitations

Evaluating Information Refusal MIRAGE aims to evaluate the information presented in a prediction. However, there is one aspect we don't capture: refusal (e.g., "There is no information on the location of the fire in the videos."). Determining a subclaim is not in supported by or contained in a large multimodal is a challenging task requiring either strong apriori knowledge of a collection or perfect retrieval and grounding within said collection. Additionally, claims refusing to answer are out-of-distribution for our entailment judges.

Multi-Video Inference Constraints As mentioned, the inability to provide multi-video inference limits the capabilities of each method presented. Extreme reduction of video frame rates to enable multi-video inference limits the performance of the VLM verifier and performing only single video judgments limits the ability to assess any cross-source inferences made by the generation system. Training VLMs for multi-video inference and optimizing multimodal inference are much needed future work for more holistic (multi-video) and scalable (optimization) evaluation.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Samuel Barham, Orion Weller, Michelle Yuan, Kenton Murray, Mahsa Yarmohammadi, Zhengping Jiang, Siddharth Vashishtha, Alexander Martin, Anqi Liu, Aaron Steven White, Jordan Boyd-Graber, and Benjamin Van Durme. 2023. Megawika: Millions of reports and their sources across 50 diverse languages. *Preprint*, arXiv:2307.07049.

Mahesh Bhosale, Abdul Wasi, Vishvesh Trivedi, Pengyu Yan, Akhil Gorugantu, and David Doermann. 2026. Craft: Critic-refined adaptive key-frame targeting for multimodal video question answering. *Preprint*, arXiv:2605.19075.

Debashish Chakraborty, Dengjia Zhang, Jialiang Jin, Hanting Liu, Katherine Guerrerio, Hanxiang Qin, Tyler Skow, Alexander Martin, Reno Kriz, and Benjamin Van Durme. 2026. Marquis: A three-stage pipeline for video retrieval-augmented generation. *Preprint*, arXiv:2605.17640.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. 2025. WavRAG: Audio-integrated retrieval augmented generation for spoken dialogue models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12505–12523, Vienna, Austria. Association for Computational Linguistics.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *Preprint*, arXiv:2411.04952.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, and Bhiksha Raj. 2024. Audio entailment: Assessing deductive reasoning for audio understanding. *Preprint*, arXiv:2407.18062.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

William Gantt, Alexander Martin, Pavlo Kuchmiichuk, and Aaron Steven White. 2024. Event-keyed summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7333–7345, Miami, Florida, USA. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pages 6465–6488, Singapore. Association for Computational Linguistics.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in LLM fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of trec 2024 biomedical generative retrieval (biogen) track. *Preprint*, arXiv:2411.18069.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4354–4374, Miami, Florida, USA. Association for Computational Linguistics.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. 2025. Videorag: Retrieval-augmented generation over video corpus. *Preprint*, arXiv:2501.05874.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. 2025. Conformal linguistic calibration: Trading-off between factuality and specificity. *Preprint*, arXiv:2502.19110.

Zhengping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khashabi, Anqi Liu, and Benjamin Van Durme. 2024. Core: Robust factual precision with informative subclaim identification. *Preprint*, arXiv:2407.03572.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063, Miami, Florida, USA. Association for Computational Linguistics.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Kelly Van Ochten, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Ronald Colaianni, Nolan King, Eugene Yang, and Benjamin Van Durme. 2025. Multivent 2.0: A massive multilingual benchmark for event-centric video retrieval. *Preprint*, arXiv:2410.11619.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2024. Overview of the trec 2023 neuclir track. *Preprint*, arXiv:2404.08071.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. Videochat-flash: Hierarchical compression for long-context video modeling. *Preprint*, arXiv:2501.00574.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2024. Videoxum: Cross-modal visual and textural summarization of videos. *Preprint*, arXiv:2303.12060.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020. Violin: A large-scale dataset for video-and-language inference. *Preprint*, arXiv:2003.11618.

Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy Lin. 2025. Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality. *Preprint*, arXiv:2505.02466.

Alexander Martin, Reno Kriz, William Gantt Walden, Kate Sanders, Hannah Recknor, Eugene Yang, Francis Ferraro, and Benjamin Van Durme. 2025. Wikivideo: Article generation from multiple videos. *Preprint*, arXiv:2504.00939.

James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W. Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, Kate Sanders, Marc Mason, and Noah Hibbler. 2024. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1904–1915, New York, NY, USA. Association for Computing Machinery.

Do June Min, Karel Mundnich, Andy Lapastora, Erfan Soltanmohammadi, Srikanth Ronanki, and Kyu Han. 2025. Speech retrieval-augmented generation without automatic speech recognition. *Preprint*, arXiv:2412.16500.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M. de Melo, Benjamin Van Durme, and Rama Chellappa. 2025. Video-colbert: Contextualized late interaction for text-to-video retrieval. *Preprint*, arXiv:2503.19009.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. Videorag: Retrieval-augmented generation with extreme long-context videos. *Preprint*, arXiv:2502.01549.

Saron Samuel, Dan DeGenaro, Jimena Guallar-Blasco, Kate Sanders, Seun Eisape, Arun Reddy, Alexander Martin, Andrew Yates, Eugene Yang, Cameron Carpenter, David Etter, Efsun Kayi, Matthew Wiesner, Kenton Murray, and Reno Kriz. 2025. Mmmorrf: Multimodal multilingual modularized reciprocal rank fusion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 4004–4009, New York, NY, USA. Association for Computing Machinery.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Ellen M. Voorhees. 2004. Overview of the trec 2003 question answering track. In *Text Retrieval Conference*.

William Walden, Marc Mason, Orion Weller, Laura Dietz, Hannah Recknor, Bryan Li, Gabrielle Kaili-May Liu, Yu Hou, James Mayfield, and Eugene Yang. 2025a. Auto-argue: Llm-based report generation evaluation. *Preprint*, arXiv:2509.26184.

William Walden, Kathryn Ricci, Miriam Wanner, Zhengping Jiang, Chandler May, Rongkun Zhou, and Benjamin Van Durme. 2025b. How grounded is wikipedia? a study on structured evidential support and retrieval. *Preprint*, arXiv:2506.12637.

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2024a. Dndscore: Decontextualization and decomposition for factuality verification in long-form text generation. *Preprint*, arXiv:2412.13175.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024b. A closer look at claim decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Pengyu Yan, Akhil Gorugantu, Mahesh Bhosale, Abdul Wasi, Vishvesh Trivedi, and David Doermann. 2026. Trace: Evidence grounding-guided multi-video event understanding and claim generation. *Preprint*, arXiv:2605.16740.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Dengjia Zhang, Alexander Martin, William Jurayj, Kenton Murray, Benjamin Van Durme, and Reno Kriz. 2026. Unified multimodal uncertain inference. *Preprint*, arXiv:2604.08701.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

A Experimental Setup Details

In this section, we outline in more detail the experiments in section 5 and section 6.

A.1 Claim Grounding Setup

In section 5, we explored how to ground claims in multimodal data. In this section, we breakdown how we setup the data, ran prediction models, and the model hyperparameters.

A.1.1 Data

In our evaluation, we report the claim verification performance on four modality combinations: text, audio, video, and audiovisual.

Text Only. For text data, we evaluate the performance of claim verification on UNLI (Chen et al., 2020) and PeopleProfile (Walden et al., 2025b). UNLI is a standard natural language inference task between a premise and hypothesis in text. PeopleProfile is a dataset from an exploration into the groundedness of a Wikipedia article in the cited sources. These annotations contain scalar support judgments (-1.0 to 1.0) from GPT-4o-mini between a subclaim and source article. This data, although synthetic, provides a verification task more salient to our evaluation setting. For both datasets, we binarize the labels, treating scores above 0.5 as supported and below 0.5 as not supported.

Audio Only. For the audio data, we evaluate the performance of claim verification on CLOTHO (Drossos et al., 2020) and the claims from WIKIVIDEO that are **only** audio support. The CLOTHO dataset has been modified by Deshmukh et al. (2024) as an audio-NLI task between an audio premise and text hypothesis with contradictory, neutral, and entailment values. We take only the contradictory (not supported) and entailment (supported) instances from the data. We leave WIKIVIDEO audio claims unmodified.

Video only. For video only, we take the WIKIVIDEO claims annotated as **only** video and/or OCR support. Other Video-NLI data exists in the literature (Liu et al., 2020), however, the underlying videos are no longer distributed. We use the WIKIVIDEO video claims unmodified.

Audiovisual. For audiovisual, we again take the WIKIVIDEO claims, but this time we take any combination of modalities, leaving the single-modality support for their respective modalities. We use the WIKIVIDEO-audiovisual claims unmodified.

Hyperparameter	Value
GPU Memory Utilization	0.9
Max New Tokens	4,096
Batch Size	4
FPS	0.5
Max Pixels	256×256
Min Pixels	256×256
Temperature	0.7
Top-p	0.95

Table 6: Model Hyperparameters

In the WIKIVIDEO data, there are claims that are unsupported by any modality. We mix these in (providing a balanced 50/50 label distribution) for the unsupported judgments with unique, randomly selected claims for each modality instance.

A.1.2 Models

We evaluate models spanning four capability profiles: audio-only (Audio Flamingo 3 (AF) (Goel et al., 2025), Qwen2-Audio (Q2-A) (Chu et al., 2024)), language-only (Qwen3 (Q3) (Yang et al., 2025)), vision-language (Qwen2.5-VL (Q2.5-VL) (Bai et al., 2025), Qwen3.5 (Q3.5) (Team, 2026)), and omnimodal (Qwen2.5-Omni (Q2.5-O) (Xu et al., 2025), CLUE (Zhang et al., 2026)).

A.1.3 Hyperparameters

The hyperparameters used for each model are listed in Table 6.

A.2 Human Alignment

In section 6, we explored how human and automatic metrics correlate when evaluating predictions. In this section, we breakdown how we setup and generated predictions.

MAGMaR Shared Task The MAGMaR shared task releases all their submissions and a set of human evaluations. Their human evaluations are a scalar score for the quality of a prediction and an overall "best" system selection. We use their scalar score as we also use a scalar score for the WIKIVIDEO annotations.

WIKIVIDEO For WIKIVIDEO, we generate predictions in three CAG (Martin et al., 2025) settings across 10 topics.

1. **LLM Only.** We generate a prediction without any information from the videos. Then, every sentence is used to retrieve one video

To help you make more accurate and consistent judgments, here is an expanded explanation of how to interpret and assign support percentages for audio-based evidence. These examples cover a range of real-world cases you may encounter in this annotation task.
100% - Full and unambiguous support: The audio clearly and directly contains the exact event, statement, or content described in the claim. There is no need for guessing or interpretation — the claim is fully verified by the audio.
80-100% - Almost complete support: The main content of the claim is clearly supported by the audio, though there may be minor ambiguity in speaker identity, context, or completeness (e.g., partial recording, mild noise). Overall, the claim is strongly supported.
60-80% - Strong partial support: The audio strongly suggests that the claim is true, but some key details may be missing, unclear, or ambiguous — such as incomplete phrases, background noise, or partial conversations. The evidence is strong but not definitive.
40-60% - Moderate partial support: There is some alignment between the audio and the claim, but large portions are missing, unclear, or open to interpretation. While the recording points in the same general direction as the claim, it lacks clarity or completeness for confident verification.
20-40% - Minimal weak support: There are small verbal cues or contextual hints that could relate to the claim, but they are insufficient to provide confidence in its truth.
0-20% - Very weak or speculative support: There may be the slightest indirect reference (such as a related topic or similar voice), but nothing concrete that verifies the claim.
0% - No support or contradiction: The audio does not relate to the claim at all, or it directly contradicts it.
Based on the provided audio and text, evaluate the probability that the text statement is true.
Your answer must be a decimal number between 0 and 1, and you must strictly follow the format below:
<answer>probability_value</answer>
Where probability_value is the result you calculate.
The text to evaluate is: {text}

Figure 3: Zero-Shot scaler prompt for audio examples.

To help you make more accurate and consistent judgments, here is an expanded explanation of how to interpret and assign support percentages. These examples are designed to cover a range of real-world cases you may encounter in the annotation task.
100% - /Full and unambiguous support: The video clearly shows the exact event described in the claim. There is no need for guessing or interpretation.
80-100% - Almost complete support: The main content in the claim is shown, but there may be minor ambiguity in location, identity, or completeness. The overall claims are supported by the video.
60-80% - Strong partial support: The video strongly suggests the claim is true, but some critical details may be missing, obscured, or ambiguous, limiting the ability to confirm the claim with certainty. The video gives strong but not definitive support.
40-60% - Moderate partial support: There is some alignment with the claim, but large portions are either missing, unclear, or open to interpretation. While the footage may point in the same general direction as the claim, it lacks the clarity or completeness needed for confident verification.
20-40% - Minimal weak support: There are small visual or audio cues that could hint at the claim, but they are insufficient to be confident.
0-20% - Very weak or speculative support: There may be the slightest indirect reference, such as a related object or setting, but nothing concrete happens.
0% - No support or contradiction: The video does not relate to the claim at all, or it directly shows something opposite.
Based on the provided video and text, evaluate the probability that the text is true.
Your answer must be a decimal number between 0 and 1, and you must strictly follow the format below:
<answer>probability_value</answer>
Where probability_value is the result you calculate.
The text to evaluate is: {text}

Figure 4: Zero-Shot scaler prompt for vision examples.

```
To help you make more accurate and consistent judgments, here is an expanded explanation of how to
interpret and assign support percentages based on textual evidence. These examples are designed to
cover a range of logic and linguistic relationships you may encounter.
100% - Full and unambiguous support (Entailment): The sentence explicitly states the information in
the claim, or the claim is a direct paraphrase of the sentence. There is no need for guessing; the
facts are identical.
80-100% - Almost complete support: The main assertions in the claim are present in the sentence.
There may be minor differences in wording, synonyms, or omission of non-essential details, but the
core meaning is fully preserved and supported.
60-80% - Strong partial support (Strong Implication): The sentence strongly implies the claim is true
through logical inference or context, though it may not state it explicitly. A reasonable person
would conclude the claim is likely true based on the sentence.
40-60% - Moderate partial support: There is a topical alignment or shared keywords. The sentence
discusses the same subject matter, but the specific assertion in the claim is neither confirmed nor
denied. It is plausible but lacks definitive evidence in the text.
20-40% - Minimal weak support: There are weak textual links, such as matching entity names or a
general theme, but the specific context is different. The sentence provides very little basis to
deduce the claim.
0-20% - Very weak or speculative support: There may be a very distant connection (e.g., related
vocabulary), but inferring the claim from the sentence would be highly speculative.
0% - No support or contradiction: The sentence is completely unrelated to the claim, or it directly
contradicts the claim (proves it false).
Based on the provided sentence and claim, evaluate the probability that the claim is supported by
the sentence.
Your answer must be a decimal number between 0 and 1, and you must strictly follow the format below:
<answer>probability_value</answer>
Where probability_value is the result you calculate.
Sentence: {sentence}
Claim: {claim}
```

Figure 5: Zero-Shot scaler prompt for text examples.

as a citation from the oracle video set using OmniEmbed (Ma et al., 2025).

2. **Oracle Video.** We generate a prediction with the oracle video set. For sentences without citations, we retrieve a video from the oracle video collection using Video-ColBERT (Reddy et al., 2025).

3. **RAG.** We first retrieve 5 videos from MultiVENT2.0 (Kriz et al., 2025) with MM-MORRF (Samuel et al., 2025). Then we generate a prediction with the set of videos in the same way as oracle generation.

B MIRAGE ++

In the main text, we present INFOF1 and CITEF1 in their standard formulations. Here, we provide variations for single-source inference and weighted evaluation, and discuss the tradeoffs between reference and collection precision.

B.1 INFOF1

B.1.1 Single-Source Precision

Single-source precision is the more computationally feasible counterpart to the multi-source formulation. Instead of scoring a claim against all evidence, we compare the claim individually against each item in the evidence and take the maximum score:

$$\text{precision}(C_P, R) = \frac{1}{|C_P|} \sum_{c \in C_P} \max_{d \in R} \mathtt{S}(c, d) \quad (6)$$

where d denotes any source of information. While single-source precision is more computationally feasible, it cannot evaluate any cross-source inferences made by a prediction system.

B.1.2 Weighted Variations

We use a notion of *information importance* to weight both precision and recall. For **precision**, information importance should be defined by the author of the prediction. For example, a human could annotate a ranked list of claims from most to least important, or a calibrated model (e.g., Jiang et al., 2025) could express confidence in the predicted claims as their importance. For **recall**, information importance should be defined by the evaluation data or the consumer of the information—in practice, annotating each reference claim for its importance to the queried information need.

We define I_i as the importance for a claim c_i. Weighted precision is:

$$\text{precision}(C_P, R) = \frac{1}{|C_P|} \sum_{c_i \in C_P} \mathtt{S}(c_i, R) \cdot I_i \quad (7)$$

For single-source inference:

$$\text{precision}(C_P, R) = \frac{1}{|C_P|} \sum_{c_i \in C_P} \max_{d \in R} \mathsf{s}(c_i, d) \cdot I_i \tag{8}$$

Weighted recall is:

$$\text{recall}(C_R, P) = \frac{1}{|C_R|} \sum_{c_i \in C_R} \mathsf{s}(c_i, P) \cdot I_i \tag{9}$$

B.1.3 Implementation Details

We support two verification backends. For the CLUE verifier, we use the scalar prompts (Figures 3, 4, 5) which return a continuous support score between 0 and 1, selecting the appropriate prompt based on document modality. With the scalar predictions we use a 0.5 threshold for supported/not supported. For the zero-shot VLM verifier, we use the binary yes/no prompts for video (Figure 6) and text (Figure 7) documents, and the subclaim-level prompt (Figure 8) for Reference Precision. In both cases, Collection Precision verifies each predicted subclaim against each document individually using single-source precision and takes the maximum score. Information Coverage uses the same prompts with the roles of prediction and reference reversed.

B.2 CITEF1

B.2.1 Single-Source Precision

As with INFOF1, we reformulate citation precision for single-source inference. Instead of scoring a subclaim against all of its cited sources simultaneously, we compare the subclaim against each cited source individually and take the maximum:

$$\text{precision}(C_P, D) = \frac{1}{|C_P|} \sum_{c_i \in C_P} \max_{d \in D_i} \mathsf{s}(c_i, d) \tag{10}$$

The set of documents D_i can either be the documents associated with the first mention of c_i, like ARGUE, or any mention of c_i depending on the desired evaluation setting. Claim de-duplication is recommended as shown by Jiang et al. (2024).

B.2.2 Weighted Variations

For CITEF1, we use a notion of *source quality* to weight both precision and recall. The quality of a source could be annotated in the evaluation data for two cases: (1) differentiating between counterfactual sources (e.g., AIGC or misinformation), and (2) human preferences—for example, a human might prefer that a system cite the most informative sources so that fewer sources need to be consulted to validate a prediction or a human might prefer a type of source (e.g., news content over social media).

We define Q_i as the quality weight for a source. Weighted citation precision is:

$$\text{precision}(C_P, D) = \frac{1}{|C_P|} \sum_{c_i \in C_P} \mathsf{s}(c_i, D_i) \cdot Q_i \tag{11}$$

For single-source inference:

$$\text{precision}(C_P, D) = \frac{1}{|C_P|} \sum_{c_i \in C_P} \max_{d \in D_i} \mathsf{s}(c_i, d) \cdot Q_d \tag{12}$$

Weighted citation recall is:

$$\text{recall}(C_R, P, D) = \frac{1}{|C_R|} \sum_{c_i^R \in C_R} \max_{d_j \in D_i} \left(\max_{s_k \in S_j} \mathsf{s}(c_i^R, s_k) \cdot Q_j \right) \tag{13}$$

B.2.3 Implementation Details

As with INFOF1, we support both CLUE and zero-shot VLM verification backends, using the same scalar prompts (Figures 3, 4, 5) and binary prompts (Figure 6, Figure 7, Figure 8) respectively.

B.3 Reference vs. Collection Precision

We discuss the tradeoffs between reference and collection precision in INFOF1 and CITEF1. The primary motivation for reference-based evaluation in multimodal RAG is computational efficiency, but each variant introduces distinct side effects.

INFOF1. Reference precision evaluates predicted subclaims against the reference instead of against the documents related to the topic (collection precision). However, human-annotated references are inherently non-exhaustive, leading to the artifact that reference precision penalizes factual claims deemed not salient to the topic. This may be a limitation if the goal is to evaluate *only* factuality: without judging subclaims against the information sources, the evaluation cannot capture the factuality of all claims in the prediction. We do not view this as a limitation for general evaluation, because it is important that information be relevant to the query and information need.

CITEF1. Reference precision for CITEF1 evaluates predicted subclaims against the reference subclaims grounded in the same cited document.

Like INFOF1, human-annotated references are non-exhaustive, so reference precision penalizes factual claims not deemed salient to the source topic. Unlike INFOF1, this **is a limitation** of reference precision: CITEF1 is designed to evaluate grounding in the cited sources, so it should not matter whether the information is relevant to the topic—only that it is supported by the attributed sources.

C Additional Human Alignment Results

In Table 7, we report additional results of Qwen3.5 9B evaluating MIRAGE on WIKIVIDEO.

D TextRAG Implementations

D.1 ALCE

In subsection 3.1, we introduce ALCE. In this section, we outline the key implementation details of this metric.

Correctness In our work, we only evaluate Claim Recall from ALCE Correctness. We do not adapt exact match against the reference because this is not suitable for long-form generation. When evaluating open-ended RAG on ELI5 (Fan et al., 2019), Gao et al. (2023) present Claim Recall, an NLI-based metric between subclaims in the reference (gold answers) and the prediction (model output). This adapts to our setting with no modification and we use the scoring function from INFOR.

Citation Quality Gao et al. (2023) evaluate citation quality with an F1 metric of citation recall and precision. When evaluating citations, Gao et al. (2023) compare a sentence from the prediction to the concatenation of cited sources. This poses two main issues in evaluation: (1) from a factuality perspective, evaluating at the sentence level requires verifying multiple subclaims in a single inference step (Min et al., 2023), and (2) from a multimodal perspective, concatenating multiple videos raises issues with computational constraints and the non-existence of multiple video training data for VLMs (Li et al., 2025; Martin et al., 2025).

To faithfully implement citation quality, we wrap verification in a loop which downsamples the framerate of the video(s) until it fits on GPU memory. We first attempt to verify a (subclaim, videos) pair at 1 fps per video. If an out-of-memory (OOM) exception is thrown, the videos are downsampled by reducing the framerate of all videos by half ($\frac{\text{fps}}{2}$) until 10 iterations of downsampling or until no OOM is reached. We never reach the 10 iterations on the videos in WIKIVIDEO, but allow for setting downsampling iterations to be defined by the evaluator. To verify the claims we use the prompt in Figure 9.

D.2 ARGUE

ARGUE (Mayfield et al., 2024) is a RAG evaluation framework developed for *report generation*—the task of producing a long-form, citation-attributed response to a complex user query. Under ARGUE, reports accumulate a series of rewards and penalties based on sentence-level judgments about whether each sentence is supported by its associated citations and whether it correctly answers any of a predetermined set of questions (or *nuggets*) associated with the report topic. ARGUE also handles various nuances related to these two dimensions, including whether a sentence *reiterates* information stated previously in the report (and thus does not require a citation) or whether a statement correctly asserts that some nugget question is unanswerable from the underlying collection. Our implementation of this metric is based largely on AUTO-ARGUE (Walden et al., 2025a).

Nuggets To evaluate predictions, ARGUE has a knowledge base of question-answer pairs on the topic called nuggets. None of our evaluation data contains nuggets, so we first we create a set of nuggets to use in evaluation. To do this, we convert our reference claims to nuggets by generating questions with an LLM (Qwen3; Yang et al., 2025) and then pairing the questions and claim as a question-answer pair.

Sentence Support is the proportion of predicted sentences attested by each of their citations. The open-source implementation of ARGUE (Walden et al., 2025a) we use implements this for a single citation at a time, which we follow. The only modification we make to this code is changing their verification of the sentence against a text document with the verification of the sentence against a video.

Nugget Recall is the proportion of nugget questions correctly answered by the report, aggregated over sentences. This is similar to ALCE Claim Recall or our INFOF1 Recall, but instead recalling QA pairs. Like ALCE, this score requires no modification for multimodal RAG evaluation. In our implementation of this, we do not use our scoring function with CLUE. Instead, we use the LLM critic from the implementation.

System Prompt: You are an expert in evaluating and verifying claims. You will be given a video, a claim, and the context the claim came from. Your task is to determine if the claim is supported by the video. You will output <response>yes<response> if the claim is supported by the video, or <response>no<response> if the claim is not supported by the video.
User Prompt: [VIDEO_HERE] Here is the context the claim came from: <claim_context> [PUT_CONTEXT_HERE] <claim_context>. Here is the claim: <claim> [PUT_CLAIM_HERE] <claim>. Only respond with <response>yes<response> or <response>no<response>. Is the claim: [PUT_CLAIM_HERE], supported by the video?

Figure 6: Zero-Shot Prompt for Claim Verification in Videos

System Prompt: You are an expert in evaluating and verifying claims. You will be given a passage of text, a claim, and the context the claim came from. Your task is to determine if the claim is supported by the passage of text. You will output <response>yes<response> if the claim is supported by the passage, or <response>no<response> if the claim is not supported by the passage.
User Prompt: Here is the passage: <verification_context> [PUT_VERIFICATION_CONTEXT_HERE] <verification_context>. Here is the context the claim came from: <claim_context> [PUT_CONTEXT_HERE] <claim_context>. Here is the claim: <claim> [PUT_CLAIM_HERE] <claim> Only respond with <response>yes<response> or <response>no<response>. Is the claim: [PUT_CLAIM_HERE], supported by the passage?

Figure 7: Zero-Shot Prompt for Claim Verification in Text

System Prompt: You are an expert in evaluating and verifying claims. You will be given a claim, the context the claim came from, and a list of claims to verify the claim against. Your task is to determine if the claim is supported by the list of claims. You will output <response>yes<response> if the claim is supported by the list of claims, or <response>no<response> if the claim is not supported by the list of claims.
User Prompt: Here is the list of claims to verify against: <verification_context> [PUT_VERIFICATION_CONTEXT_HERE] <verification_context>. Here is the context the claim came from: <claim_context> [PUT_CONTEXT_HERE] <claim_context>. Here is the claim: <claim> [PUT_CLAIM_HERE] <claim> Only respond with <response>yes<response> or <response>no<response>. Is the claim: [PUT_CLAIM_HERE], supported by the list of claims to verify against?

Figure 8: Zero-Shot Prompt for Claim Verification in Citation (Text)

Eval	Ano	INFOP		INFOR	CITEP		CITER
		R	C	R	R	C	R
EQJ	1	42.7	40.0	42.5	-21.2	-45.7	23.2
	2	34.5	40.0	14.8	-6.3	-19.7	56.2
	5	39.3	40.0	1.2	-52.7	-34.8	46.2
ICJ	1	10.0	0.0	18.5	1.8	-38.2	79.7
	2	20.0	20.0	1.8	-6.7	-53.3	62.8
	3	20.0	20.0	8.5	-6.7	-26.7	69.7
GJ	-	40.0	80.0	-18.2	-26.7	6.7	-16.3

Table 7: Kendall's Tau ($\times 100$) on WikiVideo between MIRAGE (Qwen 3.5 9B verifier) and human judgments. R: Reference, C: Collection. EQJ: Extrinsic Quality Judgment, ICJ: Intrinsic Claim Judgment, GJ: Grounding Judgment.

System Prompt: You are an expert at verifying information. You will be given a set of videos and a sentence. Your task is to determine if the sentence is fully supported by the videos. You will output <response>yes<response> if the sentence is fully supported by the videos, or <response>no<response> if the sentence is not fully supported by the videos.
User Prompt: [VIDEOS_HERE] Sentence: [PUT_SENTENCE_HERE] Is the sentence fully supported by the videos? Only respond with <response>yes<response> or <response>no<response>."

Figure 9: Zero-Shot Prompt for ALCE Citation Quality

D.3 RAGAs

RAGAs (Es et al., 2024) is a RAG evaluation framework developed for the automatic evaluation of retrieval-augmented generation—the task of producing long-form, citation-attributed responses to a user query. RAGAs consists of three main components: Faithfulness, Answer Relevance, and Context Relevance.

Faithfulness evaluates statements from a prediction against the context used in generation. We interpret these statements as subclaims to resolve the factuality issues of ARGUE and ALCE. Additionally, like Citation Quality from ALCE, this experiences the same issue of computational infeasibility in multi-source inference. However, unlike ALCE, Faithfulness is not designed such that it is necessary to perform inference on multiple sources. We modify this to verify one subclaim in one document at a time, but only using the videos used in generation and not all relevant videos. We use the same scoring function as INFOP/CITEP.

Answer Relevance evaluates the embedding similarity between the query used during prediction and an LLM-generated query based on the prediction. This metric remains unchanged from Es et al. (2024) because the generation and comparison of potential queries is text-based regardless of the information source modality. We compute this as the cosine similarity between the embeddings of the predicted and actual queries, using the Qwen3 Embedding model (Zhang et al., 2025).

Context Relevance extracts sentences from the context based on the query and computes relevance as $\frac{\text{num extracted sentences}}{\text{total number of sentences}}$. We attempt to implement this with a VLM, extracting all information relevant to the query (numerator) and extracting all information from the video (denominator). However, extracting all information from the video is challenging, as the number of potential claims is near infinite. We instead elicit a detailed summary from a VLM for the denominator. We extract information related to the query with Figure 10 (numerator) and detailed video information with Figure 11 (denominator).

In the RAGAs implementation, Es et al. (2024) give the LLM the ability to back off and say that a source has "insufficient information." We notice in our implementation that WIKIVIDEO is challenging for VLMs to make high-level inferences about and thus leads to a large number of videos deemed "insufficient information," even though humans have annotated and grounded relevant claims to the query topic in those same videos.

E Collection of Human Judgments

We collect three human judgments.

1. **Extrinsic Quality Judgments (EQJs)**. Human annotators are given information about the topic and are asked to give a 1-5 likert score considering the following attributes (in order of importance): factuality, adequacy, coherence, relevancy, and fluency.

2. **Intrinsic Claim Judgments (ICJs)**. Human annotators are given the prediction, predicted subclaims, reference, and reference subclaims, and are asked to annotate whether or not a claim is supported by the other text. This directly mirrors INFOF1-Ref: the scores are what INFOF1 would produce if a human performed the verification instead of a model.

3. **Grounding Judgments (GJs)**. Human annotators are given all the relevant videos for a topic and the predicted subclaims and asked to annotate whether or not the subclaims are supported by (grounded in) the videos. This directly mirrors CITEP: the scores are what CITEP would produce with human verification.

For each grounding judgment, we calculate the ranking between the three prediction settings and we use the scipystats (Virtanen et al., 2020) implementation of Kendall's Tau (Kendall, 1938).

E.1 Extrinsic Quality Judgments

In Figure 12, we provide the annotation instructions for the extrinsic quality judgments. For these judgments, we recruit three annotators who are native/fluent English speakers and provide them each the 10 evaluation instances and annotation instructions for the task. These annotations are redundant. We adapt these judgments from (Gantt et al., 2024).

E.2 Intrinsic Quality Judgments

In Figure 13, we provide the annotation instructions for intrinsic quality judgments. For these judgments, we recruit three annotators who are native/fluent English speakers and provide them each the 10 evaluation instances and annotation instructions for the task. These annotations are redundant.

```
Please extract relevant sentences from the provided context that can potentially help answer the
following question. If no relevant sentences are found, or if you believe the question cannot be
answered from the given context, return the phrase 'Insufficient Information'. The question is:
[PUT_QUESTION_HERE]
```

Figure 10: Zero-shot prompt for extracting query-relevant information used in the numerator of RAGAS Context Relevance.

```
Describe the video in detail and extract all the information possible from it. This includes
transcribing any on screen text (OCR) and describing any visual information beyond the summary.
```

Figure 11: Zero-shot prompt for extracting the detailed summary used in the denominator of RAGAS Context Relevance.

E.3 Grounding Judgments

We use the same annotation instructions from Martin et al. (2025) and refer the read to their work for the annotation instructions. For these judgments, we recruit 5 annotators who are native/fluent English speakers and provide them the annotation instructions and the annotation data. These annotations are not redundant.

E.4 Human MIRAGE Granular Alignment

In Table 8 we provide the alignment between each component (Precision, Recall) of human annotated INFOF1 and extrinsic quality judgments.

F Acknowledgment of AI

AI assistants were used in this project for coding and to edit text for fluency and typos.

	ICJ 1			ICJ 2			ICJ 3		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
EQJ 1	49.0	37.8	40.8	65.3	68.7	59.0	60.5	36.0	52.3
EQJ 2	61.3	49.7	68.0	76.3	53.0	74.5	81.2	51.5	81.2
EQJ 3	54.5	43.3	64.2	70.8	57.8	77.2	72.7	49.7	74.2

Table 8: Kendall's Tau agreement between MIRAGE human annotated judgments (ICJ[1-3]) and human annotated quality judgments (EQJ[1-3]).

For each prediction, you should give a likert score from 1 to 5. The criteria for giving a score is as follows (in descending importance):

- **Consistency/Factuality:** Does the article make only true statements about the topic in question, given what the reference says about that topic?
  - Articles that make factual errors, omissions, or hallucinations of any kind should be penalized.
- **Adequacy:** Does the article adequately capture all of the information contained in the reference article?
  - Articles that omit details about any of the important aspects of the topic should be penalized.
- **Coherence:** Does the article make sense on its own, as a standalone description of the topic?
  - Articles that require you to go read the reference in order to understand what they mean (or that don't make sense even then) should be penalized.
  - Articles that don't provide substance on the event should also be penalized (e.g. "the [event] happened at [a time] in [a place]")
- **Relevancy:** Does the article include only information that is relevant to the topic in question?
  - Articles that include irrelevant or superfluous information, or information about some topic other than the one represented by the reference article, should be penalized.
- **Fluency:** Does the article sound reasonable natural (like something a native English speaker might actually write)?
  - Articles that are disfluent or that sound unnatural should be penalized.

Oftentimes, some of the summaries may be very similar to each other. It is totally fine to give multiple summaries the same score if you think they are of comparable quality!
You should enter your score for each summary in the score field. The default value for each summary is 0, meaning you have not yet annotated the likert judgment. Please do not use half scores (1.5, 2.5, etc).

Figure 12: Instructions for the human likert judgments (EQJ)

In this task, you'll be given a set of 3 json files for each prediction and 3 json files for the reference. The reference json files will all be the same, but each will correspond to your annotations for a different system's predictions. In each json file, there will be the same 10 topics.
For every system prediction json, you'll see

1. **prediction:** the predicted article for the topic. This prediction is the model's attempt to write an article for the topic.

2. **claims:** this is a sub dictionary that contains the claims and sentences they came from it is formatted below. The judgment is whether the claim is supported by the reference article or not.

   { "sentence": { "claim": "the claim to be verified", "judgment": [False|True] }, }

Annotation Task
The goal of this task is to mirror how we evaluate multimodal RAG in InfoF1. The basic premise is as follows.

- For every claim in the prediction, check if it is supported by the reference article. If it is mark the judgment as True, otherwise mark it as False. (The default value is None)

- For every claim in the reference, check if it is supported by the prediction article. If it is mark the judgment as True, otherwise mark it as False. (The default value is None)

Claim Annotation Criteria
When deciding whether a **claim is supported** by the other article (reference or prediction), apply the following criteria. The goal is to make judgments **only when there is no reasonable doubt** that the claim is supported by the other text.
A claim is **SUPPORTED (True)** if:

1. **All factual elements** in the claim are explicitly stated or can be directly inferred from the other article without needing external knowledge.

   - Example: Claim: "The Eiffel Tower is located in Paris." → Supported if the other article states "The Eiffel Tower is in Paris."

2. The **meaning and intent** of the claim are **fully consistent** with the information in the article.

   - Minor wording differences are acceptable if they don't change meaning.

3. The **temporal or causal context** matches (e.g., dates, events, outcomes are consistent).

4. If the claim includes **quantitative or categorical facts** (numbers, names, locations, affiliations), these details must exactly match what is stated in the other article.

A claim is **NOT SUPPORTED (False)** if:

1. The other article **contradicts** any part of the claim.

2. The other article **omits or is ambiguous** about key details needed to verify the claim.

3. The claim requires **inference beyond what's stated**, such as outside knowledge, assumptions, or general reasoning not grounded in the text.

4. The claim is **partially supported**, but not fully — i.e., some parts are correct while others are missing or uncertain.

General rule:

Only mark a claim as **True** (supported) if you can clearly point to a sentence or set of sentences in the other article that fully confirm it, leaving **no doubt** about its accuracy. Otherwise, mark it as **False** (not supported).

Figure 13: Instructions for the human metric judgments (ICJ)