International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 11 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

A Survey of Source-Grounded Multi-Format Content Transformation: Large Language Models, Multimodal Processing, and Evaluation

Ajinkya Rasal[1], Arti Bhise[2]

[12]Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune.

Abstract: *Large language models (LLMs) make it possible to turn one source into different communication outputs like executive summaries, social media posts, operational advisories, presentations, infographics and video scripts. Creating fluent outputs in different formats doesn't mean the original facts are preserved. A trustworthy system must keep details like names, numbers, dates, conditions, uncertainties, warnings and relationships while changing the structure and tone for each format. This survey looks at research that supports source-grounded transformation across formats. It brings together work from summarization, retrieval-augmented generation (RAG) multimodal understanding, document intelligence, speech recognition, video understanding, controllable generation and factuality evaluation. The system break the process into an end-to-end pipeline: source ingestion, modality-specific extraction, canonical evidence representation, grounded planning, format-specific generation and verification. The review shows a problem: most studies focus on improving one modality or one output task while cross-format consistency and traceability across different deliverables are not well tested. The research suggest evaluation criteria that include source coverage, factual support, contradiction rate, semantic preservation, format compliance, readability, latency and cost. This survey offers a research taxonomy, compares methods, highlights open problems and proposes a reference architecture. It is a draft literature synthesis. Screening and eligibility decisions based on specific databases still need to be completed before submission, to a journal.*

Index Terms: Content transformation, factual consistency, large language models, multimodal AI, retrieval-augmented generation, summarization, document intelligence, speech recognition.

I. INTRODUCTION

Organizations get information in different forms: reports, policy papers, scanned pages photos, recordings and videos. The same information might need to go to groups using different ways to share it. A manager might want a summary an operations team might need something they can act on and the public might get a short message for social media. Large language models and models that work with types of data can help with parts of this process but a result made by a model can sound good and convincing while changing the original information a little.

Its called as source-grounded -format content transformation when one or more original items are turned into several results with clear rules. Every part of each result should be able to be linked to the original source or to other clear evidence. The goal is not just to have language; it is to change the information in a controlled way while keeping the main meaning the same.

This problem touches on areas like summarizing text retrieving information models that work with more than one type of data analyzing documents, understanding speech and videos making things that can be controlled and checking if information is true. These fields have parts but the goals for checking them are not the same. This review adds: (1) a way to describe the task; (2) a way to group the different parts; (3) a look, at how different studies work; and (4) what is missing and what needs to be checked for systems that make several results that all agree with each other and use the same information.

II. RESEARCH METHODOLOGY

This paper uses a narrative survey along with a structured way to search and screen for information. The main time period covered is from 2018 to 2026. Some older studies are kept if they define a method or concept that is widely used today. Six research questions direct the review:

- RQ1: What techniques take information and turn it into short or specific text?
- RQ2: How do documents, images, audio and video get turned into formats that generative systems can use?
- RQ3: How do large language models and RAG systems make sure the content they create is based on sources?
- RQ4: What methods check if the content is accurate, honest and keeps the meaning?
- RQ5: What stops systems from keeping the same quality when they produce different formats?
- RQ6: What kinds of structures and tests could help solve these problems?

The search includes places like IEEE Xplore, ACM Digital Library, ACL Anthology, SpringerLink, ScienceDirect, PMLR, NeurIPS proceedings and arXiv. The search terms are grouped into four parts: task terms like content transformation, summarization, data-to-text and controlled generation; grounding terms like RAG, factual consistency and attribution; modality terms like document AI, OCR, vision-language, speech recognition and video understanding; and evaluation terms like factuality semantic preservation, benchmark and cross-modal consistency. An example search is: (LLM OR language model). Grounded generation OR factual consistency OR RAG) AND (summarization OR content generation OR multimodal).

Studies are included if they suggest, look into or test a method that helps with source extraction, representation, retrieval or grounding controlled generation -modal understanding or evaluating the output. Journal and conference papers that have been checked by others are given priority. Some preprints are added if they are important or have technical details and are marked as such. Studies not related to the topic duplicate versions, materials not in English and those with unclear information are not included.

The first search and checking of sources for this version were done on 10 November 2026 using publisher and academic pages. The selected studies are grouped by the type of work they do. This is not a PRISMA review yet: exact numbers, from the database, how many were removed how many were excluded after checking and how much reviewers agreed have not been recorded. Before sending it in the database files and screening notes should be. The final numbers shown in a diagram. This way the work is not shown as a review when it is still an early search.

III. FOUNDATIONS AND ENABLING TECHNOLOGIES

A. Text Generation and Summarization

Sequence-to-sequence pretraining made generation better. BART learns by fixing text and works well when fine-tuned for summarization and other generation tasks [1]. PEGASUS uses gap-sentence generation picking sentences to rebuild and was tested across many summarization areas [2]. These techniques help compress and rephrase text effectively. Just because a summary overlaps with the original does not mean every claim, in it must follow from the input.

B. Retrieval-Augmented and Evidence-Grounded Generation

RAG mixes generation with non-parametric retrieved evidence.

In the RAG design RAG gives retrieved passages to a generator for tasks that need a lot of knowledge. Later studies look at ways to do retrieval, augmentation and generation. These ways include retrieval, reranking, query rewriting and iterative retrieval. In systems that transform text retrieval can find source passages that back up every claim, in the output. A source-only mode is still needed when external knowledge is not allowed.

C. Multimodal Representation and Understanding

In my view multimodal learning covers representation, alignment, fusion, translation and co-learning [5]. CLIP aligns textual representations using large-scale image–text supervision [14]. BLIP-2 links frozen image encoders and language models, through a querying transformer [15]. LLaVA on the hand looks into visual instruction tuning [16]. All of these approaches help interpret images but a practical workflow still has to separate what text is actually visible what descriptions are. What observations remain uncertain.

D. Document Intelligence

Documentsre structured visual objects. They include things like tables, headings, the order in which you read them areas on a page and elements that stand out. LayoutLMv3 works by combining text and image information. It uses a way to mask parts of the data and aligns words with image patches. This model helps with understanding forms answering questions, about documents and analyzing layouts [6]. This approach encourages a way of processing documents that keeps track of positions and the structure of turning a PDF or a scan into just plain text.

E. Speech and Video Processing

Speech recognition turns sound into text. Mistakes in transcribing can carry over into summaries and alerts. The research find that Whisper uses a big weak-supervision approach to make speech recognition stronger. Video-LLaMA mixes sound encoders with a language-model interface to read video. A real-world process might pick out frames write down speech, spot scene changes and match times. Evidence must keep time marks so that claims can be followed back, to the parts of the video.

F. Controllable Generation

The results of a transformation depend on who's reading it how its written, how long it is, what language it uses where its shared and why its created. Research on text generation that can be controlled looks at how to manage content and features, by adjusting models using learning techniques giving instructions working with parts of the system and making changes when the text is being created [17]. These controls should only affect how something is presented, not the facts. The style directions should always follow the rules of what's true.

IV. TAXONOMY AND REFERENCE PIPELINE

The systems are sorted based on five factors: input modality, evidence representation, grounding strategy, output family and verification method. This classification system shows where information might get lost during processing.

Dimension	Categories	Design question
Input	Text, PDF/scan, image, audio, video	What evidence can be extracted?
Representation	Text, layout blocks, transcript, frames, timestamps	What provenance is retained?
Grounding	Source-only, RAG, hybrid	Can each claim link to evidence?
Output	Summary, post, advisory, infographic, slides, script	Which facts must remain invariant?
Verification	Rules, NLI, QA, alignment, human review	How are unsupported claims detected?

A reference pipeline has seven steps: ingestion and validation; second modality-specific extraction; third canonical evidence construction; fourth task and audience planning; fifth parallel artifact generation; sixth claim-level verification; and seventh packaging with provenance. The canonical evidence representation is central. It may include facts, source spans, entities, quantities, dates, conditions, modality references, confidence and extraction metadata. Output generators use this shared representation of reinterpreting raw source. The system must distinguish extracted facts, derived statements that use explicit reasoning and external context that comes from approved sources. Each type must have a provenance label. When the source is unclear or missing the system should keep uncertainty. Ask for clarification not invent details. Multi-format generation is therefore a constrained rendering problem, over shared evidence, not a set of prompts.

V. COMPARATIVE ANALYSIS OF REPRESENTATIVE RESEARCH

Table II lists studies by how they help the pipeline. This is a comparison, not a ranking: the studies use different tasks, datasets and metrics so the scores they report cannot be compared directly.

Studies	Contribution	Pipeline role	Limitation for multi-format task
[1], [2]	Sequence-to-sequence summarization	Text transformation	Summary scores do not prove claim support.
[3], [4]	RAG and retrieval design	Evidence grounding	Retrieved context can be irrelevant or misused.
[5], [14]–[16]	Multimodal/visual-language models	Image understanding	Does not ensure cross-output consistency.
[6]	Layout-aware document model	PDF/document ingestion	Extraction varies by layout and language.
[7], [8]	Speech and video understanding	Audio/video ingestion	Transcription and temporal errors propagate.
[9]–[13]	Factuality and faithfulness metrics	Verification	Mostly evaluate one response against one context.

A. Synthesis of Research Gaps

I first notice that modality pipelines are often tested on their own. OCR, speech transcription and video understanding each have benchmarks yet downstream systems must measure how errors in extraction change the final products. Second factuality evaluation usually looks at one answer against one context. A system that writes a summary, an advisory, a post and a slide deck makes a set of outputs that should agree with one another and with the source. Third format compliance and factuality are often split: a short post may skip a warning while a slide might alter a number. Fourth provenance is not reliably shown. Finally latency and cost rise when many outputs are made at once and verification calls, on models.

VI. SOURCE GROUNDING AND EVALUATION

A. Factuality and Faithfulness

I first notice that modality pipelines are often tested on their own. OCR, speech transcription and video understanding

each have benchmarks yet downstream systems must measure how errors in extraction change the final products. Second factuality evaluation usually looks at one answer against one context. A system that writes a summary, an advisory, a post and a slide deck makes a set of outputs that should agree with one another and with the source. Third format compliance and factuality are often split: a short post may skip a warning while a slide might alter a number. Fourth provenance is not reliably shown. Finally latency and cost rise when many outputs are made at once and verification calls, on models.

B. Proposed Evaluation Framework

- Source coverage means how much of the facts from the source stay in the final product.
- Claim support is the share of claims that are tied to a portion of the source text or a trusted outside reference.
- Contradiction rate shows how many claims go against the source or against generated pieces.
- Semantic preservation means keeping the names, numbers, dates, places, conditions, negative words and unsure parts exactly as they appear.
- Format compliance checks that the output follows the rules for length, format, target readers and the fields that are needed.
- Readability and usefulness are how clear, actionable and suitable for the reader the text feels.
- Operational performance covers how long it takes, how tokens or compute power it uses how often it fails and how much extra checking is needed.

A good benchmark must have papers that have tables, dates, limits, special rules and conditional steps; pictures that have OCR; recordings that have different accents and background sounds; and videos where what is said does not match what is shown. Every item must have a list of facts that come from the source and the limits, on the output. Problematic examples should show numbers on different pages, unclear dates, instructions that are turned off and filenames that trick the user.

VII. CHALLENGES AND OPEN PROBLEMS

Evidence extraction errors occur when OCR, speech recognition, or frame sampling fails to capture information accurately or alters important facts. Therefore, confidence estimates should be passed to subsequent processing stages to indicate the reliability of extracted information. Long-context reasoning presents another challenge, as relevant information may be distributed across multiple pages or timestamps. Extracting isolated portions can result in the loss of relationships between statements, such as a warning that is later modified by an exception. Cross-format consistency is also important because different generated outputs may vary in detail. Content compression must therefore preserve essential facts and contextual information. Furthermore, grounding under style pressure is a significant concern, as persuasive formats may encourage exaggeration or unsupported claims. Style controls must not compromise the accuracy of source evidence. Multilingual and domain variations introduce additional challenges because translation and specialized terminology may alter the intended meaning. Finally, human oversight is essential in high-impact workflows and should include risk-based reviews, source references, uncertainty indicators, version histories, tool records, and mechanisms for correcting errors.

VIII. FUTURE RESEARCH DIRECTIONS

A promising direction is a provenance-aware representation. In a provenance-aware representation every tiny claim keeps track of its source span, modality, timestamp or page location confidence and how it has changed. Another helpful idea is constrained multi-output generation. With constrained multi-output generation the research make artifacts from one verified evidence graph and then check that all of those artifacts agree. Risk-sensitive verification must focus on numbers, deadlines, safety instructions and legal duties. Benchmark development should test sets of artifacts that come from the source instead of single isolated summaries. Interactive systems should show citations. Let people look at or fix the evidence, behind each statement. For a platform converting reports, PDFs, images, audio and video into summaries, social posts, advisories, infographics, presentations and video packages these findings suggest an architecture. A modular architecture uses modality-specific extractors, an evidence store, a planner, specialized renderers and a verifier. This is a design implication taken from the reviewed literature, not a result that was experimentally proven in this survey.

IX. PROBLEM FORMULATION AND SYSTEM REQUIREMENTS

A. Formalizing the Transformation Task

Let S be a collection of source items. Each item can be text, a document page, an image, an audio segment or a video interval. Let A be the requested artifacts. Each artifact has a specification that includes its audience, channel, language, length, tone, structure and purpose. A transformation system creates A by using S and C. C is the set of constraints. This system is different from generation. The goal is to make sure the content is useful and easy to read. The content must

follow the evidence constraints. Every factual claim must be supported by S or, by a source that is allowed.

B. Functional and Non-functional Requirements

The task involves functional requirements. First ingestion has to find out the file type, the encoding how long the file's how many pages it has and whether extraction is possible before it starts generation. Second extraction must keep details such as page numbers, table cells speaker turns and timestamps. Third the system needs to build a shared evidence representation that every requested output can use. Fourth the planner must turn specifications into content requirements. Fifth generation must keep the facts while changing the layout and the words. Finally verification must spot claims, contradictions, missing parts and format problems, before the final delivery.

B. Functional and Non-functional Requirements (continued)

Non-functional requirements are just as important as ones.

Traceability means that users must be able to follow a claim to where it came from.

Robustness means that the system must deal nicely with files that are broken audio that's missing pictures that are low quality or transcripts that are not complete.

Scalability means that the system must keep memory use and token use in check when it deals with long documents or videos.

Privacy means that the system must handle sensitive source material carefully and follow rules about how it keeps that material.

Usability means that any uncertainty or failures, in extracting data should be shown clearly not hidden inside text.

These requirements show that quality must be judged at the level of the system not only by looking at the language model's output

C. Invariants and Permitted Transformations

The properties that must remain consistent across generated artifacts include names, dates, quantities, units, locations, causal qualifications, negations, warnings, and expressions of uncertainty. In contrast, presentation-related properties such as sentence order, vocabulary level, degree of detail, rhetorical framing, and visual arrangement may be adapted according to the target format. For example, when generating a social media post, the system may omit background explanations while preserving safety conditions that affect the meaning of an instruction. Similarly, when converting content into a presentation, paragraphs may be represented as charts or other visual elements; however, the corresponding values and labels must remain correctly aligned. This distinction between preserving factual information and adapting its presentation is a fundamental principle of source-grounded content transformation.

X. MULTIMODAL INGESTION AND EVIDENCE NORMALIZATION

A. Text and Document Processing

Plain text is the simplest form of input; however, even plain text files may require line-break adjustment, encoding correction, and the removal of repeated headers, footnotes, and section-boundary artifacts. Processing PDF documents introduces additional challenges because PDFs may contain selectable text, scanned images, vector diagrams, or a combination of these elements. Therefore, a reliable document-processing pipeline should identify the document type and select an appropriate extraction method, such as direct text extraction or Optical Character Recognition (OCR). Layout-aware models, such as LayoutLMv3 [6], highlight the importance of preserving document coordinates and visual context during processing. Tables should be represented as structured rows and columns rather than being merged into unstructured text strings. Furthermore, page and section identifiers should be retained so that generated claims can be traced back to their original locations in the source document.

B. Image and Diagram Processing

Images can have photos, screenshots, charts, maps, labels or documents inside them. A vision-language model can describe what it sees. The description is not the same as an actual measurement. For example when looking at a chart the interpretation must keep the labels the units, the legend items and any uncertainty in reading the values. Text that comes from optical character recognition (OCR) should be clearly different from what the model guesses based on the image. If a diagram shows connections the system should include nodes, edges and labels much as possible. Models like CLIP [14] BLIP-2 [15] and visual instruction tuning [16] help make this possible. Still the system must keep track of which parts of the response came from text, in the image and which came from interpreting the elements.

C. Audio Processing

Audio ingestion usually means decoding audio detecting when someone is speaking, recognizing speech separating speakers and lining up timestamps. Whisper [7] shows how useful large-scale supervised speech recognition can be but the confidence, in a transcription can change when accents, noise, overlapping speakers or special vocabulary are present. Therefore a transcript should keep timestamps. If possible show who is speaking. Numbers, names and negations need checking because a tiny mistake can change an advisory or executive summary a lot. Also sounds that are not speech can matter in some fields so a system should not assume that every audio source is fully captured by its transcript.

D. Video Processing

Video combines information over time with an optional audio track. A working system can pick frames at scene changes or at time intervals. It can also pull out words find text that appears on the screen and line up all these pieces on the same timeline. Taking frames at fixed times is easy to do. Might skip quick events. Using scene changes to pick frames can cut down on repeated content. It needs a good way to find where one scene ends and another begins. Video-LLaMA [8] shows how audio and visual data can be used with language models. When transforming information based on sources every piece of data must keep its timestamp and the type of input it came from. For example saying "the speaker demonstrates a procedure" should be backed by the actual time period, in the video where this happens. A spoken idea should be matched to the part of the transcript not guessed from a file name or any other indirect clue.

E. Canonical Evidence Representation

After extracting information in a modality- way the system must normalize the data into a standard evidence format. A good evidence record includes: (i) a proposition, (ii) a source identifier, (iii) a location such as a page number, region, timestamp or transcript offset (iv) the modality used, (v) the extracted text or visual observation (vi) confidence level or review status and (vii) connections, to entities, quantities and other propositions. The format should keep track of where the information came from not store a long prompt string. This ensures the evidence can be reused across types of artifacts and allows a verifier to check the origin of any single claim.

XI. PLANNING AND GENERATING MULTIPLE ARTIFACTS

A. Artifact Specification

I believe each output type should be described by a specification instead of only a free-form prompt. An executive summary may need a decision-oriented overview, key findings, risks and recommended actions. A LinkedIn post may need an opening, a professional tone and a concise narrative. An operational advisory may need scope, affected parties, action steps, deadlines and escalation instructions. An infographic may need labels and a visual hierarchy. A presentation may need slide-level structure and speaker notes. A video package may need a script a scene plan, voice-over and on-screen text. These specifications should be. Testable.

B. Shared Planning Before Parallel Generation

Generating all items on your own from the source can lead to different ways of understanding the same thing. A better way is to create a shared content plan. This plan includes facts, supporting proof, what the audience cares about what should not be assumed and what is missing. Each person who makes the product then chooses and rearranges the proof based on what their product needs. Making things at the time can save time but it should only start after the proof and plan are ready. This keeps the question "what do we know?" separate, from "how should the system share it?". Lowers the chance that different ways of asking for results create conflicting facts.

C. Cross-artifact Consistency

Consistency is not about using the words or giving the same level of detail. It is about making sure that shared claims do not conflict. If the executive summary says the deadline is 18:00, then a social media post cannot suggest 20:00. If an advisory names a certain district as impacted an infographic must not show an area as affected unless there is proof. A consistency checker can break down each piece of content into individual claims. It can look at things, like entities, numbers, dates, tone and conditions. When there is a contradiction it should be settled using the source not by seeing what most of the outputs say. The source is always the authority.

D. Human Review and Interaction

Human review should be risk-based. Human review can let low-risk stylistic choices pass automatically. Human review must give stronger checks to safety instructions, financial

values, legal obligations, medical statements and emergency deadlines. The interface should show the evidence behind claims, spot low-confidence extraction and let people correct the main evidence record. When a reviewer changes a fact all dependent artifacts should be marked stale and rebuilt or rechecked. This feedback loop is far better, than editing one output and leaving versions inconsistent.

XII. COMPARATIVE DISCUSSION AND DESIGN TRADE-OFFS

A. Source-only Grounding versus Retrieval-Augmented Generation

Source-only grounding works best when the transformation needs to rely on the material that is provided like internal reports or incident advisories. In these cases no outside information should be added. Retrieval-augmented generation or RAG can bring in context, definitions or up-to-date details. But this comes with a risk—external facts might mix with the source facts, leading to confusion or incorrect results. To avoid this the system must clearly label where each piece of evidence comes from. It should also let users turn off retrieval when needed. If retrieval is turned on any passages that are pulled in should be cited separately. They should not. Hide conflicting information that is already in the source. RAG [3] [4] offers a framework for this setup. Still the final decision on whether external evidence's allowed depends on governance rules. These rules vary based on the task, at hand.

B. End-to-end Models versus Modular Pipelines

End-to-end multimodal models can reduce the complexity of integration. Might capture relationships that exist between different types of data. However they can make it hard to tell if an error came from perception from retrieval from reasoning or, from generation. Modular pipelines offer checkpoints and let specialized OCR, ASR, document parsing and verification parts be replaced on their own. Their problems include the complexity of managing the flow, delays and errors that can spread from one step to another. A hybrid architecture is therefore appealing: multimodal models help with interpretation while explicit evidence records and clear checks keep track of facts.

C. Automatic Metrics versus Human Evaluation

ROUGE [22] looks at how much words overlap between the summary and the reference. It is still useful in some cases when comparing summaries.. It cannot check if the summary is factually correct. Pyramid evaluation [23] focuses on content units. It does a job of showing how well the system selects important information.. It needs people to label the data, which takes time and effort.

Then there are methods like FactCC [9] QAGS [10] AlignScore [11] FActScore [12] and ARES [13]. These give kinds of automatic signals. Each one has strengths.. None of them is perfect. They should not be used in place of one another. Instead they should be used together to get a picture.

For systems that produce outputs the evaluation needs to go beyond just looking at each output alone. It should also check if there are contradictions, between the outputs. It should look at how the sources are covered.. It should make sure key information is not lost during compression.

D. Quality, Latency, and Cost

Multi-artifact workflow can multiply model calls: extraction, planning generation, for each output claim extraction, verification and repair. The research find that multi-artifact workflow that generates outputs in reduces wall-clock time yet it raises concurrent resource demand. Cost-aware system can share extraction and evidence construction reuse representations and give stronger verification only to high-risk claims. Caching helps when the same source and artifact specification recur, long as source versioning stops stale outputs. Evaluation should report quality with latency, token use, failure rate and human-review effort.

XIII. BENCHMARK DESIGN AND REPRODUCIBLE EVALUATION

A. Benchmark Unit and Dataset Composition

The natural benchmark unit is not a summary; the natural benchmark unit is a source bundle that comes with a set of requested artifacts. Each source bundle should contain one or more source modalities and an annotated inventory of facts.

The benchmark should cover PDFs and noisy PDFs, multi-page tables, scans, screenshots, charts, speech, with varied accents videos that have both spoken words and visual events and collections that mix different modalities. The benchmark should also be ready to handle mixed-modality collections.

Domain diversity is important because a system that works well on news articles may fail on instructions or technical reports.

B. Annotation Protocol

Annotators must find facts note where they come from and rate their importance. They should also assess uncertainty. Identify relationships like cause, condition or exception. Every artifact created must be checked for supported claims, unsupported claims, contradictions, missing facts and whether it follows the required format. If there are disagreements a second round of annotation can help resolve them. The protocol needs to say what counts as a material claim and be able to tell the difference between a harmless rewording and a change that alters meaning. Instructions, examples and rules for resolving disputes should be made public along, with the benchmark.

C. Recommended Metrics

A reproducible evaluation should report least seven dimensions. Source coverage measures how well the evaluation retains the required facts. Claim support measures the share of claims that are linked to valid evidence. Critical-fact recall measures whether the evaluation keeps the important facts. Contradiction rate measures how often the evaluation conflicts with the source or related artifacts. Provenance completeness measures whether the evaluation can point to usable source locations, for each claim. Format compliance measures how well the evaluation follows the required constraints. Operational measures include latency, cost and failure rate. Human ratings of clarity, usefulness and audience suitability complement these evaluation metrics. Results should be split by modality, type and risk level.

D. Error Analysis and Ablation

Aggregate scores hide problems. In error analysis errors are sorted into extraction, alignment, retrieval, reasoning, generation, formatting or verification failures. Ablation studies remove timestamps, layout structure, evidence confidence, shared planning or cross-artifact checking to see how much each part matters. Robustness tests involve misleading filenames, source passages, negated statements, unit conversions, low-quality OCR overlapping speakers and short but important video events. These robustness tests check if the system stays safe when the evidence is unclear.

XIV. PRACTICAL ARCHITECTURE IMPLICATIONS

A. Reference Architecture

A deployable platform can be organized into six services: an ingestion gateway; modality-specific extractors; an evidence normalization service; an artifact planner; specialized generators; and a verification and delivery service. The ingestion gateway validates file type and size. Extractors produce text, layout, image observations, transcripts, and timestamped frame descriptions. The normalization service stores evidence records. The planner prepares a common content plan and artifact specifications. Generators create the requested deliverables. The verifier checks source support and cross-artifact consistency before delivery.

B. Failure Handling

Failures need to be clear and specific. A video that does not have a stream should not be shown as successfully transcribed; it might still be handled visually if the task allows. A scanned document that has OCR should be labeled as low confidence and sent for checking. If a file is too big for processing the system should tell the user about the limit. Suggest a proper solution, like dividing the file. Partial extraction should not be considered as coverage of the source. These steps are important because a clean result can hide an input process.

C. Security, Privacy, and Governance

Source documents can hold commercial or other confidential information. A production platform must set access controls, transport security, retention and deletion policies logging practices and rules for sending data to model providers. The user must know if external retrieval is on and which sources can be used. Audit logs must capture source versions, model and prompt versions, artifact specifications and verification outcomes. Governance is not separate, from factuality: if reliable records are missing a reviewer cannot rebuild why an artifact had a claim.

XV. SYNTHESIS OF THE LITERATURE BY PIPELINE STAGE

A. What Existing Research Solves Well

The reviewed literature gives us building blocks for each part of the process. Neural summarization gives us ways to shorten and rewrite text [1] [2]. Retrieval-augmented generation helps include evidence when making new content [3] [4]. Multimodal learning and vision-language models allow us to understand both text and images together [5] [14]–[16]. Document models keep track of layout. How things are arranged on a page [6]. Research on speech and video helps us pull information from moving images [7] [8]. Factuality research gives us tools both built with people, in mind to check if what is generated is true or supported [9]–[13]. All these parts together make it possible to build a system that can transform content across formats.

B. What Remains Fragmented

The main problem is that these capabilities are usually made and checked using task definitions. A summarization model is usually checked by looking at one text output. A document model is checked by looking at extraction or question answering. An ASR system is checked by looking at transcription error. A RAG system is checked by looking at retrieval or answer quality. These checks do not directly answer whether one source bundle can be turned into artifacts without losing essential facts. They also do not show whether a claim, in a post can be traced to a specific table cell, image region or video interval. The gap is not just that there is no generator but that there is no integrated evidence and evaluation layer.

C. Implications for Research Design

Future studies should compare workflows under common source bundles and artifact specifications. For example a single incident report could be used to create a summary, a public notice, an operational checklist and a slide outline. The evaluation should measure not the quality of each artifact but also whether the outputs agree on the affected location, deadline, severity and required action. A second experiment could change the ingestion method while keeping the underlying facts the same, which would let researchers see how OCR, ASR and video extraction errors affect the results. This design would make it possible to compare architectures, on the end-to-end goal.

XVI. TAXONOMY OF FAILURE MODES AND MITIGATION

A. Extraction and Alignment Failures

Extraction failures happen before generation. Optical character recognition might leave out a minus sign, a point, a table heading or a footnote. Speech recognition can mix up a noun or flip the order of a number. Frame sampling may skip a important event. Alignment failures occur when the extracted information is linked to the page speaker, timestamp, table column or visual object. To reduce these issues keep the source, store location metadata set confidence thresholds use validation rules that fit the domain and send uncertain observations that could have a big impact, to human review.

B. Evidence Selection and Reasoning Failures

System may extract facts but choose only a subset that changes overall meaning. Warning may be separated from the condition that limits it; reported increase may be presented without its baseline;. Exception may be omitted from a short output. Reasoning failures include causal claims, incorrect arithmetic and treating correlation, as causation. Mitigation requires importance labels, relationship aware evidence records and a requirement that generated claims identify their supporting propositions. Verifier should test conditions and qualifiers well as entity and number matching.

C. Generation and Presentation Failures

I see that generation failures bring hallucinated details, exaggerated certainty, unsupported recommendations and style-driven embellishment. The research also notice that presentation failures occur when a correct fact is placed under the heading paired with an incorrect chart label or shortened until its condition is lost. When dealing with artifacts, text and graphic elements should be validated together: a chart can contain correct values but misleading axes and a slide can show a correct number beside an unrelated category. It is observed that artifact-specific validators should therefore check both content and structural relationships.

D. Verification and Governance Failures

The verifier may itself produce assurances especially when the verifier relies on the same model family and prompt assumptions as the generator. The proposed approach suggests that verification should rely on multiple signals. These signals include checks for dates and quantities source-span matching, independent model judgments and human review, for high-risk claims. Governance failures show up when source versions are missing prompt changes are not. Manual edits go unrecorded. For reproducibility the system need a record that lists the source, extraction version, evidence representation, generation configuration, verification result and final artifact revision.

XVII. LIMITATIONS OF THIS SURVEY

A. Scope Limitations

I read this paper. The paper see that it is a focused narrative survey with a structured search protocol. This paper is not a systematic review or meta-analysis. This paper connects research areas that use datasets, evaluation protocols and definitions of quality. Because of that this paper makes an functional comparison; this paper does not rank models or compare reported scores as if they were measured under a common benchmark. The references chosen by this paper are representative of enabling research, than an exhaustive inventory of every relevant system.

B. Evidence and Reproducibility Limitations

The search protocol described in Section II identifies databases, query concepts, eligibility criteria and screening principles. However the search protocol does not contain a screening log that shows exact hit counts, duplicate counts, full-text exclusion reasons and reviewer agreement. The reference list and bibliographic metadata should be compared with publisher records before the paper is submitted. A publication-ready systematic version of the search protocol would need dated database exports reproducible query strings, a documented screening process and a flow diagram. These limitations are stated explicitly so that the paper does not claim a level of coverage that the search protocol has not yet shown.

C. Generalization Limitations

The proposed pipeline and evaluation framework are recommendations based on the research that was reviewed. They have not been tested through a controlled experiment in this survey. Performance may change depending on the language used the domain, the quality of the source material the model chosen the length of the context and the level of risk involved in the task. Especially automatic factuality metrics may not catch mistakes or missing information, in tables, images or time-based media. Because of this the framework should be seen as a research agenda and a design proposal that needs to be tested through experiments and data.

XVIII. CONCLUSION

Here is a survey that links summarization, RAG, multimodal learning, document intelligence, speech and video understanding, controllable generation and factuality evaluation with the goal of source-grounded multi-format content transformation. The research literature gives building blocks yet gaps still exist in claim-level provenance, cross-format consistency and collection-level evaluation. A trustworthy system must first separate extraction from generation then make evidence visible change style while keeping facts intact and finally check every piece against the evidence. Moving forward benchmarks are needed that test not if a single answer sounds good or is true but also if a whole set of deliverables, for a particular audience is complete, consistent and traceable. Benchmarks must evaluate summarization, RAG, multimodal learning, document intelligence, speech and video understanding, generation and factuality evaluation.

REFERENCES

[1] M. Lewis et al., “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” Proc. ACL, 2020, pp. 7871–7880, doi: 10.18653/v1/2020.acl-main.703.

[2] J. Zhang, Y. Zhao, M. A. Saleh, and P. J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” Proc. ICML, PMLR, vol. 119, 2020.

[3] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Proc. NeurIPS, 2020.

[4] “Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv preprint, 2023. (Verify authors, version, and final publication metadata before submission.)

[5] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[6] Y. Huang et al., “LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking,” Proc. ACM Multimedia, 2022, pp. 4083–4091, doi: 10.1145/3503161.3548112.

[7] A. Radford et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” Proc. ICML, PMLR, vol. 162, 2022, pp. 28492–28518.

[8] H. Zhang, X. Li, and L. Bing, “Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding,” Proc. EMNLP: System Demonstrations, 2023.

[9] W. Kryscinski et al., “Evaluating the Factual Consistency of Abstractive Text Summarization,” Proc. EMNLP, 2020, pp. 9332–9346, doi: 10.18653/v1/2020.emnlp-main.750.

[10] A. Wang, K. Cho, and M. Lewis, “Asking and Answering Questions to Evaluate the Factual Consistency of Summaries,” Proc. ACL, 2020, pp. 5008–5020, doi: 10.18653/v1/2020.acl-main.450.

[11] Y. Zha et al., “AlignScore: Evaluating Factual Consistency with a Unified Alignment Function,” Proc. ACL, 2023, pp. 11328–11348, doi: 10.18653/v1/2023.acl-long.634.

[12] S. Min et al., "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation," Proc. EMNLP, 2023.

[13] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," Proc. NAACL-HLT, 2024.

[14] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," Proc. ICML, PMLR, vol. 139, pp. 8748–8763, 2021.

[15] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," Proc. ICML, PMLR, vol. 202, 2023, pp. 19730–19742.

[16] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," Proc. NeurIPS, vol. 36, 2023.

[17] X. Liang et al., "Controllable Text Generation for Large Language Models: A Survey," arXiv:2408.12599, 2024.

[18] Y. Wang et al., "Factuality of Large Language Models: A Survey," Proc. EMNLP, 2024, pp. 19519–19529, doi: 10.18653/v1/2024.emnlp-main.1088.

[19] S. Lamsiyah, A. Nourbakhsh, and C. Schommer, "Trust but Verify: A Comprehensive Survey of Faithfulness Evaluation Methods in Abstractive Text Summarization," Proc. RANLP, 2025.

[20] H. Zhang et al., "A Survey on Multimodal Large Language Models," National Science Review, vol. 11, no. 12, 2024, doi: 10.1093/nsr/nwae403.

[21] P. Wang et al., "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," arXiv:2409.12191, 2024.

[22] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," Text Summarization Branches Out, ACL Workshop, 2004, pp. 74–81.

[23] A. Nenkova and R. Passonneau, "Evaluating Content Selection in Summarization: The Pyramid Method," Proc. HLT-NAACL, 2004, pp. 145–152.

[24] Y. Liu et al., "Multilingual Denoising Pre-training for Neural Machine Translation," Trans. Assoc. Comput. Linguistics, vol. 8, pp. 726–742, 2020, doi: 10.1162/tacl_a_00343.

[25] J. Zhang et al., "Text-to-video generators: a comprehensive survey," J. Big Data, vol. 12, art. 253, 2025.