

Review

Managing the Unmanageable: Multimodal Artificial Intelligence for Unstructured Data Management and Analysis

Chong Ho Yu ^{*}, Nino Miljkovic and Zhaoyang Wang

Department of Mathematics, College of Natural and Computational Sciences, Hawaii Pacific University, 1 Aloha Tower Drive, Honolulu, HI 96813, USA; nino.miljkovic@gmail.com (N.M.); z8173720@gmail.com (Z.W.)

^{*} Correspondence: cayu@hpu.edu

Abstract

Today, data are no longer confined to numerical values arranged in row-by-column matrices or stored neatly within relational databases. One of the defining characteristics of big data is its high variety, encompassing unstructured and multimodal forms such as text, audio, images, and video. These data types dominate contemporary domains including social media, digital humanities, biomedical research, education, and surveillance systems. Yet these data types remain difficult to manage and analyze using traditional data management architectures. To cope with this shift, modern data management systems must move beyond schema-driven designs and incorporate multimodal artificial intelligence capable of understanding, integrating, and reasoning across heterogeneous data modalities. This article examines how multimodal AI, in particular large multimodal foundation models, can be leveraged to support the ingestion, representation, organization, and analysis of unstructured data. It discusses emerging multimodal data management frameworks, outlines a conceptual pipeline for multimodal data analysis, and highlights key challenges related to scalability, interpretability, and governance. By situating multimodal AI at the core of data management, this work argues that effective data analysis in the era of big data requires systems that treat meaning, context, and cross-modal relationships as first-class computational objects rather than afterthoughts.

Keywords: multimodal AI; RAG; LLM; personal intelligence; neuro-symbolic AI; world model; unstructured data



Academic Editors: Jorge Bernardino,
Le Gruenwald, Elio Masciari and
Laurent D'Orazio

Received: 10 July 2026

Revised: 11 August 2026

Accepted: 14 August 2026

Published: 17 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

1.1. Research Problem

Traditionally, data have been equated with numerical representations, reflecting a realist assumption that reality can be objectively measured, quantified, and modeled [1,2]. Qualitative researchers have long challenged this view, arguing that many complex human phenomena, such as lived experience, meaning-making, consciousness, and social interaction, cannot be adequately reduced to numbers without significant loss of depth and context. This critique is closely aligned with phenomenology, particularly the work of Edmund Husserl, who argued that modern science had become detached from the lifeworld. Specifically, this refers to the pre-theoretical world of everyday experience that gives phenomena their meaning [3]. From this perspective, numerical data may describe how much or how often, but they often fail to capture how it is experienced. Later qualitative methodologists extended this insight, emphasizing that human realities are socially constructed, context-dependent, and interpretive, requiring methods such as in-depth

interviews, participant observation, and textual analysis to access meanings that resist quantification [4,5]. Rather than rejecting numbers outright, qualitative scholars argue for epistemological pluralism. This idea states that numeric data are powerful for certain questions, but understanding complex phenomena (identity, suffering, belief, learning, or ethical judgment) demands approaches that honor subjectivity, intentionality, and lived experience alongside measurement [6].

For much of the twentieth century, critiques of quantitative dominance were often ignored or minimized because numerical methods were widely perceived as more scientific, objective, and rigorous, aligning neatly with positivist ideals of measurement, replicability, and statistical inference [1]. However, this hierarchy has been increasingly destabilized by the explosive growth and practical importance of unstructured data. The issue stems from the fact that audio, images, video, and certain text forms do not naturally conform to tabular or numeric formats.

For example, cybersecurity now relies heavily on biometric modalities, such as facial recognition, voice authentication, and behavioral patterns, all of which originate as rich, non-numeric signals that must be interpreted rather than simply counted [7,8]. Customer support systems routinely archive massive volumes of chat transcripts, call recordings, screenshots, and photos uploaded by users, where meaning, intent, and context are embedded in language and visuals rather than in predefined variables. Beyond these examples, healthcare increasingly analyzes clinical notes, medical images, and radiology scans; autonomous vehicles depend on video streams and sensor fusion; social media analytics centers on posts, images, and videos; while legal and financial sectors mine contracts, emails, and compliance documents for patterns and risk signals [9]. These developments have forced a conceptual shift. Rigor is no longer equated solely with numerical abstraction, but with the ability to systematically analyze meaning-rich, high-dimensional, and context-dependent data. As a result, qualitative insights and interpretive methods, which are now often augmented by machine learning and AI, are being re-valued as essential components of scientifically robust inquiry rather than as inferior alternatives to quantitative analysis.

As such, the digital ecosystem is currently undergoing a significant structural transition, characterized by a shift from deterministic data management, predicated on structured, tabular information, to probabilistic systems of reasoning capable of ingesting and interpreting the chaotic reality of unstructured media [10]. For decades, the enterprise data stack was architected around the Relational Database Management System (RDBMS) and structured query language (SQL), a paradigm that, while efficient for transactional records, inherently excluded the vast majority of human-generated information. Estimates consistently indicate that approximately 80% of enterprise data exists in unstructured forms: video footage, audio recordings, complex PDF contracts, satellite imagery, and social media interactions [11].

Historically, these virtually “unmanageable” data were relegated to “data swamps” that acted as passive object storage accessible only via limited metadata or specialized computer vision pipelines that required extensive manual feature engineering. With the advance of text mining, which is based on computational linguistics and natural language processing, analysts could analyze open-ended data using tools like SAS Text Miner, JMP Pro, and IBM SPSS Modeler [2,6,12–15]. However, these approaches remain limited to textual rather than audiovisual data.

To a substantial degree, NoSQL (Not Only SQL) did alleviate the structural rigidity of the RDBMS, but it solved the problem of storage and schema rather than the problem of meaning. While the RDBMS required data to be normalized into strict rows and columns (a “schema-on-write” approach), NoSQL introduced schema-on-read, which allowed enterprises to ingest unstructured and semi-structured data, like JSON documents, sensor

logs, and social media feeds, without first redesigning a rigid database schema [16]. While NoSQL solved the structural limitation (how to store a PDF or a log file), it did not inherently solve the semantic limitation. Specifically, a NoSQL database can store a million video files or medical notes, but it cannot “read” them to understand their content, let alone analyze them.

However, this transition should not be understood as a wholesale replacement of one paradigm by another. In practice, modern data architectures are increasingly hybrid: relational databases now incorporate vector search extensions (e.g., pgvector for PostgreSQL, vector indexing in Azure SQL), while AI-native platforms retain structured metadata layers for governance and auditability. The trajectory described in this review is therefore best read as a continuum rather than a binary, with the most capable production systems combining SQL, vector search, object storage, and AI-based retrieval within unified pipelines. Section 4.4 examines this complementarity in detail.

The maturation of multimodal foundation models has fundamentally inverted this dynamic [17]. Three related terms require careful distinction. A large language model (LLM) primarily processes and generates language. A large multimodal model (LMM) processes two or more modalities, such as text, images, audio, or video. The term multimodal large language model (MLLM) is commonly used for an LMM in which a large language model serves as the central generative or reasoning component, often connected to modality-specific encoders or trained jointly on multimodal data [18]. Because terminology remains inconsistent across the literature, this review uses LMM as the broader umbrella term. The capability of LMM transforms unstructured data from a static liability into a dynamic asset, thus enabling “any-to-any” retrieval and analysis. Essentially, we are moving from “Systems of Record,” which capture what happened, to “Systems of Reasoning,” which can analyze why it happened by synthesizing insights across modalities [19].

This review provides a structured technical analysis of this transition. It examines the emerging architectural frameworks, specifically the evolution from static Retrieval-Augmented Generation (RAG) to dynamic, autonomous Agentic RAG workflows [20,21]. It compares frontier multimodal models, including Google’s Gemini, OpenAI’s GPT, and Anthropic’s Claude, across benchmarked task performance, documented architectural characteristics, and reported deployment considerations (Section 5). Furthermore, this review examines the infrastructure challenges accompanying this transition, including the scalability of vector search, the interpretability of opaque embeddings, and the governance required to manage bias and Personally Identifiable Information (PII) in probabilistic outputs [22]. Finally, use cases of Google’s Personal Intelligence are discussed to illustrate how multimodal AI can function as an ecosystem that synthesizes structured and unstructured data.

The argument of this review is organized around a single thread: The management of unstructured data has been constrained by three successive limitations. The first is a structural limitation—how to store data that do not fit rows and columns—which relational systems could not overcome and which NoSQL largely resolved. The second is a semantic limitation—how to make systems understand what stored artifacts mean—which persisted through the NoSQL era and is only now being addressed by multimodal foundation models and embedding-based retrieval. The third is an epistemic limitation—how to trust, verify, and govern the probabilistic interpretations these models produce—which emerges precisely because the semantic limitation has been overcome. The remainder of the review follows this progression.

Section 2 traces the transition from unimodal, schema-driven architectures to multimodal foundation models and introduces the shift from Systems of Record to Systems of Reasoning. Section 3 examines how this shift is operationalized through a three-stage

pipeline of ingestion, representation, and vector storage. Section 4 analyzes retrieval-augmented generation (RAG) and its agentic extensions as the mechanism that connects stored representations to model reasoning, and clarifies the continuing, complementary role of structured databases. Section 5 compares the frontier foundation models that power these pipelines. Section 6 addresses the epistemic limitation directly, covering scalability, interpretability, verification, and governance. Section 7 through Section 9 look forward: emerging paradigms, a proposed taxonomy that synthesizes the preceding sections into seven architectural dimensions, and applied use cases that illustrate the framework in research practice.

1.2. Review Methodology

This article is a structured narrative review rather than a systematic review in the PRISMA sense. Given the pace of change in multimodal AI—where foundational results frequently appear first as preprints or vendor technical reports months before peer-reviewed publication—a strictly systematic protocol would have excluded much of the field’s current state. We therefore adopted a transparent, criteria-driven narrative approach (consistent with methodological guidance for structured narrative reviews [23–26]) designed for breadth and currency, while documenting our process below.

Searches were conducted iteratively between September 2025 and June 2026, comprising an initial search phase and two targeted update phases carried out during successive rounds of peer review. Literature was retrieved through structured queries across Google Scholar, PubMed, ProQuest Dissertations & Theses Global, ResearchGate, Semantic Scholar, arXiv, and JSTOR, and supplemented with AI-assisted discovery tools, including Gemini Deep Research, Claude AI, and Scite AI, to identify emerging work. Representative query terms included “multimodal large language model,” “retrieval-augmented generation,” “vector database,” “multimodal embedding,” “agentic RAG,” and “unstructured data management,” combined with modality terms (text, image, audio, video). Reference lists of key sources were also examined (backward snowballing). Whereas the academic databases were queried with keyword and Boolean combinations, the AI-assisted discovery tools were prompted with full natural-language research questions, for example: “Please conduct deep research to examine how multimodal AI—particularly large multimodal foundation models—can be leveraged to support the ingestion, representation, organization, and analysis of unstructured data.”

Technical literature on multimodal models, RAG architectures, and vector infrastructure was generally restricted to work published in 2023–2026 to reflect the current state of a fast-moving field; earlier foundational works were retained where necessary to establish key concepts or architectures, while foundational theoretical, epistemological, and database-architecture sources were included regardless of date to ground the conceptual framing. We included English-language peer-reviewed papers, preprints, technical reports, and primary vendor documentation that directly addressed multimodal data management, representation, retrieval, or governance. We excluded sources that were purely promotional without verifiable technical detail, duplicative of stronger sources, articles published in predatory journals, or outside the review’s scope. For the RAG literature specifically, we excluded papers that merely employ an out-of-the-box vector database or a standard RAG pipeline as a baseline tool for a task in which the AI system itself is not under investigation, and included papers in which the architecture, retrieval mechanism, indexing algorithm, or multimodal alignment is itself the object of study. Because this is a structured narrative review conducted through iterative, exploratory searching rather than a protocol-driven systematic review, record-level screening counts were not logged, and a PRISMA-style flow accounting is therefore not reported; narrative reviews are not expected to provide such

accounting, and their quality is instead assessed through transparency of search strategy, inclusion logic, and source-quality criteria [23–26], which we document here. Because record-level retrieval, deduplication, and screening counts were not systematically documented during the review process, a fully reproducible account of study selection cannot be provided, and the possibility of selection bias cannot be entirely ruled out.

Because the field’s frontier is defined largely by industrial laboratories, we deliberately drew on multiple source types and treated them differently. Source quality was assessed through a five-tier evidence hierarchy: (1) peer-reviewed journal articles and archival conference proceedings, which ground the review’s core scientific and theoretical claims; (2) preprints (chiefly arXiv) and technical reports, used for findings too recent to have completed peer review and identified as such in the reference list; (3) benchmark publications and evaluation platforms, used for quantitative performance comparisons and reported with model version and evaluation date; (4) vendor documentation, model cards, and official announcements, used exclusively to describe proprietary system capabilities and labeled as vendor-reported in the text; and (5) practitioner commentary in technical blogs, used sparingly for deployment observations and labeled as indicative. Where a claim is central to the review’s argument, it is anchored in tier-1 or tier-2 sources, with lower tiers treated as supplementary. This tiered approach reflects the reality of a fast-moving field in which many developments appear first in technical reports and provider documentation, while preserving the distinction between validated evidence and vendor-stated capability.

The final corpus comprises the 101 substantive works cited in this review (excluding the four methodological references cited in this section): 20 peer-reviewed journal articles, 12 peer-reviewed conference papers, 10 books and book chapters, 23 preprints and technical reports, 19 vendor and institutional documents (including model cards and official announcements), 14 technical blog and practitioner commentaries, and 3 benchmark or evaluation platforms. Peer-reviewed publications thus constitute approximately one-third of the corpus—42% including scholarly books—a proportion that reflects the field’s current publication dynamics, in which a substantial share of frontier results has not yet completed peer review.

The retrieved corpus was organized thematically into the categories that structure this review: the unimodal-to-multimodal transition, the multimodal data pipeline (ingestion, representation, storage), retrieval-augmented generation, the comparative model landscape, governance and operational challenges, and emerging directions.

To mitigate reliance on any single source, quantitative performance claims were, wherever possible, attributed to primary benchmark sources and cross-checked against independent results; claims supported only by vendor material are identified as such in the text. Because benchmark leadership among frontier models can reverse between releases, comparative statements are framed as time-stamped snapshots rather than stable rankings.

2. From Unimodal to Multimodal

2.1. Limitations of the Unimodal Era

This section establishes the first two strands of the review’s thread: Why schema-driven architectures could eventually store unstructured data but could not understand it, and how multimodal foundation models resolve this semantic limitation. The history of data management is largely a history of enforcing structure upon chaos. The “Big Data” era of the early 2010s expanded the volume and velocity of data ingestion but struggled to address variety effectively. Traditional analytical workflows created deep silos. Text was processed by Natural Language Processing (NLP) pipelines, images by Computer Vision (CV) Convolutional Neural Networks (CNNs), and tabular data by SQL engines. These silos were rarely interoperable. For example, a sentiment analysis model processing

customer reviews had no mathematical way to “see” the image of the damaged product attached to the same support ticket [18].

This unimodal fragmentation resulted in “brittle” workflows. For example, engineering a pipeline to analyze traffic footage required training specific object detectors for cars, pedestrians, and traffic lights. If the requirement changed to detecting “erratic driving behavior,” the entire pipeline often had to be re-engineered. The cost of this feature engineering meant that only the most high-value unstructured data were ever analyzed, leaving the “long tail” of enterprise information dormant in cold storage [11].

2.2. *The Rise of the Multimodal Foundation Model*

The transition examined in this review is grounded in two established bodies of scholarship. In the machine learning literature, Baltrušaitis, Ahuja, and Morency formalized multimodal machine learning around five core technical challenges—representation, translation, alignment, fusion, and co-learning—a taxonomy that still structures the field [27]. Liang, Zadeh, and Morency subsequently reframed these challenges for the foundation-model era, emphasizing heterogeneity, cross-modal connections, and interactions as the central theoretical problems [28]. The pivotal enabler of progress on all five challenges was representation learning: the replacement of hand-engineered features with learned, distributed representations that capture the underlying explanatory factors of the data [29]. Multimodal embeddings are the direct descendants of this program—and, as this review argues, they are also the object that data management systems must now ingest, index, and govern. Recent surveys of multimodal large language models document how rapidly this representational substrate has been scaled into general-purpose perception and reasoning systems [30].

A parallel evolution occurred in the data systems literature. The recognition that variety—not merely volume or velocity—constitutes the defining analytical challenge of big data predates the foundation-model era [31], and the lakehouse architecture was proposed precisely to unify the schema-flexible storage of data lakes with the management and governance guarantees of warehouses [32]. The multimodal lakehouse discussed later in this review is best understood as the convergence of these two research traditions: representation learning supplies the semantic layer, and lakehouse architecture supplies the management layer. Framing the review this way clarifies what is genuinely new—the treatment of meaning, context, and cross-modal relationships as first-class computational objects—and what is continuous with prior data management scholarship.

The advent of the Transformer architecture and subsequent Large Multimodal Models (LMMs) introduced a unifying substrate: the embedding [33]. Unlike traditional feature vectors, which represent specific attributes (e.g., “contains_red_pixel”), multimodal embeddings represent semantics in a continuous vector space. Models like CLIP (Contrastive Language-Image Pre-training) were trained on hundreds of millions of image–text pairs to minimize the geometric distance between an image and its accurate textual description [34]. Furthermore, advanced algorithms are utilized to compress vision tokens, making multimedia storage and retrieval viable [35].

This mathematical alignment implies that in the vector space, the numerical representation of a photograph of a golden retriever is proximal to the numerical representation of the text string “dog playing in the grass.” This “Any-to-Any” capability allows for cross-modal interoperability without explicit metadata tagging. A user can search a video archive using a text query (“Find the moment the CEO gives a demo of the robot”), or search a document database using an image query (“Find contracts with this specific logo”) [18].

The implications for data management are far-reaching. The database is no longer just a storage engine; it becomes a dynamic perception engine or an engine of pattern

recognition. By integrating LMMs, databases can “watch” video and “listen” to audio, indexing the semantic content dynamically. This creates a “Multimodal Lakehouse” where raw data files and their semantic indices coexist, accessible via natural language queries rather than complex code. Specifically, a multimodal lakehouse is a unified data architecture that combines the schema-flexible storage of a data lake for raw, unstructured files with the structured querying capabilities of a data warehouse, augmented by semantic vector indices so that multimodal data and their embeddings coexist and are queryable through natural language [36].

2.3. From Systems of Record to Systems of Reasoning

The ultimate trajectory of this technology is the creation of Systems of Reasoning. A System of Record answers typical quantitative questions like “What was the total revenue in Q3?” through a deterministic lookup. A System of Reasoning answers questions like “Why did revenue drop in Q3, considering the customer sentiment in support calls and the visual evidence of supply chain disruption in our logistics video feeds?” In this case, the query goes beyond superficial numbers to qualitative insight [37]. Table 1 summarizes the conceptual shift from “Systems of Record” to “Systems of Reasoning”.

Table 1. Evolution of Data Management Paradigms.

Feature	RDBMS	Hybrid/Transitional	Multimodal AI
Data Focus	Structured (Table/Row)	Structured + semi-structured + selective unstructured	Unstructured (Video/Audio/Image/Text)
Approach	Schema-on-write	Schema-on-write + vector indices + NLP enrichment	Schema-on-read/Embedding
Query Type	Keyword/ID Lookup	SQL + semantic search (e.g., pgvector, Azure AI Search)	Semantic/Concept Search
System Goal	Systems of Record	Augmented Systems of Record	Systems of Reasoning
Function	Static Storage	Semi-Dynamic Storage	Dynamic Perception and Pattern Recognition
Inference	Deterministic	Deterministic + selective probabilistic	Probabilistic

3. The Architecture and Process of Multimodal Intelligence

Multimodal AI refers to artificial intelligence systems that jointly process, interpret, and generate information across two or more data modalities (e.g., text, images, audio, video), in contrast to unimodal systems that operate on a single modality. While Section 2 explained why the semantic limitation demands multimodal foundation models, this section examines how that capability is operationalized in practice: through a three-stage pipeline—ingestion, representation, and storage—that turns raw multimodal files into computable semantic objects. To operationalize multimodal AI, organizations must transition from ad hoc data science projects to robust, scalable engineering pipelines. The emerging reference architecture for this is the AI-Native Data Platform, which fundamentally rethinks the traditional Extract–Transform–Load (ETL) process.

3.1. Stage 1: Ingestion and the “ObjectRef” Paradigm

The first challenge in multimodal data management is the physical disconnect between analytical engines (Data Warehouses) and unstructured storage (Data Lakes/Object Storage). Historically, a data warehouse like Snowflake or BigQuery stored metadata (file-names, timestamps), while the actual .mp3, .mp4 or .pdf files resided in AWS S3, Microsoft Azure, or Google Cloud Storage. Analyzing the file content required moving data out of

the warehouse, processing it, and writing results back, which was often a slow, expensive, and insecure round-trip.

Several methods have been developed to ease this bottleneck, such as SAS's in-memory analytics [12] and edge computing [38]. A critical innovation addressing this is the ObjectRef data type, exemplified by recent updates to Google BigQuery [37]. ObjectRef allows the SQL engine to treat an unstructured file in cloud storage as a first-class citizen within a table. A table row can contain a customer ID (integer), a transaction date (timestamp), and an ObjectRef pointing to the audio file of the customer's support call.

This architectural unification enables Zero-ETL analysis. Users can run SQL queries that invoke remote functions (powered by Vertex AI or similar services) to process the referenced object directly. For instance, a query could look like:

```
SELECT customer_id, ML.TRANSCRIBE(audio_file) FROM support_calls WHERE
duration >300
```

This integrates unstructured inference directly into the structured query planner, reducing data movement and simplifying governance [37].

For organizations without the resources to build custom ingestion pipelines, managed services like Amazon Bedrock Data Automation have emerged. These services act as an intelligent ingestion layer that automatically identifies file types, parses content (e.g., extracting text from PDFs, transcribing audio), and prepares it for downstream indexing [11]. This abstracts away the complexity of managing libraries like ffmpeg or Tesseract, providing a standardized JSON schema output for every ingested asset.

3.2. Stage 2: Representation and Embedding

Once ingested, data must be converted into a format suitable for machine reasoning. This is the domain of Embeddings. The core mechanism of multimodal representation is the high-dimensional vector. Modern embedding models, such as Amazon Nova Multimodal Embeddings or Google's multimodal gecko models, are designed to project text, images, and video into a single semantic space [18]. The dimensionality of these vectors typically ranges from 768 to 4096 dimensions. Higher dimensionality captures more nuance but increases storage and search costs.

A key innovation is the move towards Cross-Modal Embeddings that do not require an intermediate text step. Early multimodal systems often captioned an image and then embedded the caption. True multimodal models embed the image directly, capturing visual features (texture, lighting, spatial composition) that text descriptions often miss [18].

One of the most complex aspects of multimodal representation is Chunking. A vector represents a single semantic unit; however, a 20-page PDF or a 1 h video cannot be compressed into a single vector without substantial information loss. Therefore, data must be segmented as follows [39]:

1. Text chunking moves beyond fixed-size windows toward semantically coherent chunking (sentence or paragraph level) and adaptive chunking, in which LLMs or "late chunking" preserve document-level context.
2. Image chunking segments images through patch-based grids, object-centric detection (e.g., YOLO), or scene-graph approaches that preserve the relationships between visual elements.
3. Audio chunking relies on silence-based segmentation or speaker-based (diarization) chunking; splitting audio in the middle of a speaker's turn blends semantic signals and degrades retrieval quality.
4. Video chunking uses shot-based and action-based detection so that temporal segments capture complete events rather than fragmented frames.

5. Cross-modal alignment is the frontier of chunking: it synchronizes disparate data types, for example, ensuring that a text description remains linked to the specific video timestamp it describes.

3.3. Stage 3: The Vector Storage Infrastructure

The rapid growth of vector data necessitates specialized storage infrastructure. For example, LanceDB offers a serverless, embedded architecture. It utilizes the Lance columnar format, which is optimized for machine learning data. Unlike traditional databases that require a running server, LanceDB can run in-process (like SQLite) and query data directly from S3. This “Multimodal Lakehouse” approach considerably simplifies the stack for developers, as the vector index and the raw data live together [36].

4. Retrieval-Augmented Generation

4.1. What Is RAG?

The pipeline described in Section 3 makes multimodal data representable; retrieval-augmented generation is what makes them amenable to reasoning. RAG is therefore the connective tissue of the review’s thread, linking stored semantic representations to the reasoning capabilities of foundation models. Retrieval-Augmented Generation (RAG) is an architectural framework in artificial intelligence that enhances the capabilities of generative foundation models by integrating an external retrieval mechanism. It has become the standard pattern for grounding LLMs. In a multimodal context, RAG involves retrieving relevant images/video clips and feeding them into the context window of an LMM [40]. While traditional foundation models generate responses based solely on the patterns learned during their training phase, RAG-enabled systems first search an external database or corpus to find relevant information related to a specific query. This retrieved content is then provided as additional context to the generative model, which synthesizes the final output. This process ensures that the AI’s responses are grounded not solely in parametric memory but also in verifiable, up-to-date, or domain-specific evidence, thereby significantly reducing the risk of ‘hallucinations’ [41–43].

A critical challenge in modern AI is the gap between raw, unstructured multimodal data (e.g., high-resolution images, audio recordings, and video streams) and the sophisticated “understanding” or reasoning capabilities of foundation models. RAG acts as the functional bridge between these two domains. By transforming raw multimodal artifacts into searchable semantic representations (often through multimodal embeddings), RAG allows foundation models to “see” and “hear” specific data points within a vast database. Instead of the model attempting to memorize every detail of a multimodal corpus, the RAG bridge selectively feeds the model only the most relevant snippets of data required for a specific reasoning task, ensuring a streamlined and context-aware inference process [44–46].

While foundation models possess extensive “internal” knowledge through a static repository of information encoded within their billions of parameters during pre-training, this knowledge is often fixed in time and lacks the granularity required for specific, data-intensive tasks. To alleviate the issue, RAG provides the necessary “external” context drawn from a multimodal database, encompassing text, audio, video, and image data. This external context grounds the AI’s reasoning in the actual data being managed. For example, in medical imaging or radiology, a model might use its internal knowledge of anatomy but rely on RAG to retrieve specific, similar cases from a multimodal database to provide a grounded, evidenced interpretation of a new scan. This distinction transforms the model from a generalist into a specialized agent capable of reasoning over private or highly dynamic datasets [45,47].

Traditionally, data management systems have functioned as passive storage units where information is retrieved using specific “row” identifiers or keyword matches. RAG fundamentally transforms these systems into active retrieval systems. By utilizing semantic vector search, RAG enables users and AI agents to query a database based on “concepts” or “semantic meaning” rather than exact string matches. This shift allows for the discovery of relationships across different modalities. For example, finding a video segment that semantically matches a textual description of an event can be done without the need for manual tagging or rigid indexing. Consequently, data management evolves from a simple archival process into a dynamic, semantic layer that actively supports the reasoning and decision-making capabilities of AI systems [48].

4.2. Limitations of Standard RAG

However, this conventional or standard RAG approach has significant limitations.

1. Context Window Bottlenecks: Although powerful AI models can handle many tokens, filling the context window with large numbers of retrieved chunks (“Context Stuffing”) can increase noise, degrade reasoning performance and increase latency/cost [49]. The “Lost in the Middle” phenomenon means models often forget information buried in the center of a large prompt [50].

2. Textual Dominance: The internal “space” where MLLMs store information is overwhelmingly biased toward text. Even when processing images, the model’s neurons focus mostly on linguistic concepts, which suppresses the fine-grained visual details necessary to distinguish one image from another during retrieval [51].

3. The Alignment Trade-off: MLLMs allocate significant computational resources toward bridging the representational gap between distinct modalities (i.e., aligning image and text embeddings). While this is great for generating captions, it homogenizes the embeddings. This means different images start to look mathematically similar to the model, reducing the “discriminative power” needed to select the exact right match from a list [51].

4.3. The Rise of Agentic RAG

To overcome these limitations, the industry is shifting toward Agentic Architectures. Agentic RAG is an advanced retrieval-augmented generation architecture in which the AI system autonomously decomposes queries into sub-tasks, routes them across multiple tools and data sources, re-ranks candidate results, and iteratively refines its retrieval strategy—rather than following a fixed, single-pass retrieve-then-generate pipeline. Specifically, Agentic RAG is characterized primarily by autonomous decision-making and goal-oriented behavior. Unlike older RAG, which follows a rigid, linear sequence of fetching and then generating, Agentic RAG acts as a reasoning-capable assistant that can dynamically plan its approach. Key characteristics include the ability to break complex queries into sub-tasks, use diverse tools like web search or specialized APIs via the Model Context Protocol (MCP), and iteratively refine its results until a sufficient answer is reached [21,52].

Naive or standard RAG paradigms rely on static, “one-shot” retrieval. If the initial search fails to find the right information, the system produces an incomplete or hallucinated response. Agentic RAG uses Planning and Routing to decompose complex queries into smaller sub-tasks. It can decide in real time whether to query a structured SQL database, a vector store, or a live web search tool based on the specific needs of the prompt [21].

To combat the “garbage in, garbage out” problem, Agentic RAG implements ReRanking. It generates multiple candidate answers from diverse contexts and uses a specialized Ranking Agent to evaluate and pick the most accurate response, rather than just relying on the top-k vector search results. Furthermore, while older systems are confined to a static knowledge base, the agentic version can identify gaps in its own information. If a

document index lacks current data, the agent can autonomously decide to use a web search tool or a live API to supplement its context. More importantly, agentic RAG introduces “reflection,” allowing the system to check its own work. If a retrieved chunk is irrelevant or a tool call fails, a Fallback Agent can switch strategies mid-stream, providing a level of flexibility that traditional, sequential pipelines cannot achieve [21,52]. Figure 1 illustrates the workflow of both the standard/simple version and the advanced version of RAG.

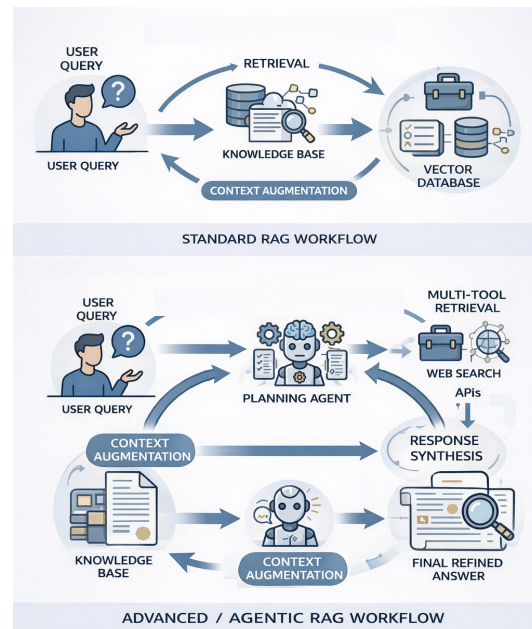


Figure 1. Standard/Naive and Advanced Agentic RAG Frameworks.

4.4. Relationship Between RAG and Conventional Databases

The growing capabilities of multimodal artificial intelligence (AI) systems and the rapid advancement of retrieval-augmented generation (RAG) architectures do not necessarily imply a diminishing role for traditional structured databases, such as SQL-based systems. On the contrary, these developments underscore the continued importance of structured data management as a foundational component of trustworthy AI systems.

Large language models (LLMs), when used in isolation, operate by generating responses based on probabilistic patterns learned from vast corpora of text. Although these models demonstrate remarkable fluency and general knowledge, they are prone to factual inaccuracies and hallucinations, particularly when responding to queries involving proprietary, domain-specific, or time-sensitive information [20]. RAG addresses this limitation by enabling LLMs to retrieve relevant information from external, authoritative knowledge sources at inference time. Retrieved passages from databases, document repositories, or knowledge bases are injected into the model’s prompt, thereby grounding the generated response in verified context rather than solely in the model’s internal parameters.

This grounding mechanism substantially improves factual accuracy and reliability, making RAG-based systems suitable for enterprise and mission-critical applications, including financial reporting, internal policy support, and clinical documentation, where creativity is undesirable and errors carry significant consequences [53]. In such contexts, the integrity of the underlying data pipeline becomes as important as the sophistication of the AI model itself.

It is at this juncture that structured data management approaches, particularly SQL-based systems, reassert their relevance. If an AI model functions as the cognitive engine of an intelligent system, the data pipeline serves as the infrastructure that ensures information

is delivered accurately, securely, and consistently. Within this pipeline, SQL plays a central role in enforcing constraints, validating transactions, and mediating access to trusted data sources. In RAG workflows, data pipelines must both prepare data through cleaning, normalization, and governance and retrieve it efficiently during inference. Modern SQL platforms increasingly support hybrid querying paradigms that combine traditional relational filters (e.g., temporal constraints, identifiers, and access controls) with semantic retrieval mechanisms based on vector embeddings [53].

Although platforms such as SQL Server continue to evolve in their native support for vector search, they are clearly moving toward tighter integration with AI workloads, often within architectures that pair relational databases with specialized vector or search services. The central issue is not whether SQL systems replace vector databases, but rather how SQL can anchor semantic retrieval within an auditable, secure, and well-governed data foundation. This anchoring is essential for maintaining transparency, compliance, and accountability in enterprise AI systems.

As multimodal AI systems become increasingly capable of interpreting images, audio, and video, the relevance of SQL does not diminish. Organizations will continue to require reliable “sources of truth” for operational, regulatory, and financial data. Domains such as healthcare, finance, and public administration demand data that are consistent, traceable, and ACID-compliant. ACID compliance refers to the guarantee that database transactions uphold atomicity, consistency, isolation, and durability, ensuring that operations remain correct, predictable, and auditable even under concurrent access and system failures [54].

While analysts may rely on multimodal models to interpret medical images or summarize customer service interactions, they will still depend on SQL databases to manage patient histories, billing records, contracts, and access permissions with precision and regulatory compliance.

As AI agents become more autonomous, the role of data professionals is likely to shift from manual query construction via SQL and other conventional tools toward higher-level system design. Practitioners will increasingly act as AI orchestrators, responsible for defining schemas, constraints, views, and retrieval strategies that guide intelligent systems toward safe, reliable, and interpretable behavior. Rather than fading into obsolescence, SQL and structured data management approaches will remain indispensable tools for imposing order, governance, and clarity in an increasingly probabilistic AI landscape.

5. The Model Landscape: A Comparative Analysis

Every stage of the architecture described in Sections 3 and 4 ultimately depends on the capability of the foundation model at its core; this section therefore compares the frontier models that currently power multimodal pipelines. The foundation model is the computational core of the multimodal pipeline. The current landscape is shaped by rapid competition among Google, OpenAI, and Anthropic, each with distinct architectural philosophies. The comparative analysis of these frontier multimodal models was organized around three explicitly defined dimensions: (i) benchmarked task performance, defined as published accuracy on standardized public benchmarks (e.g., Video-MMMU for knowledge acquisition from long-form video, and MMMU/MMMU-Pro and CharXiv for document, chart, and diagram reasoning); (ii) documented architectural characteristics, such as context-window length and native versus pipeline-based modality processing; and (iii) reported deployment considerations, such as latency, cost, and integration tooling. Throughout, we distinguish quantitative benchmark results—attributed to primary sources and reported with the model version and evaluation date—from practical deployment claims, which frequently derive from vendor documentation and technical blogs and are therefore treated as indicative rather than confirmatory. Because the foundation-model landscape shifts on a

near-weekly basis and benchmark leadership can reverse between releases, all comparative statements are time-stamped and should be read as a snapshot rather than a stable ranking.

The specific versions compared here—Google’s Gemini 3 Pro (released November 2025), OpenAI’s GPT-5.2 (released December 2025), and Anthropic’s Claude Opus 4.5 (released November 2025)—constitute the frontier generation for which directly comparable public benchmark evidence exists as of mid-2026 [55–57]. Each provider has already shipped incremental successors: Gemini 3.1 Pro Preview and Claude Opus 4.6 in February 2026, and GPT-5.5 in April 2026 [58–60]. The comparison is therefore presented, as before, as a documented snapshot that illustrates how a multi-benchmark evaluation framework operates in practice, not as a current leaderboard.

5.1. Video Analysis: Long-Form Understanding and Context Capacity

Video is among the most demanding tasks for multimodal models due to its temporal dimension and large data volume, and the evaluation instrument itself has evolved since the previous generation: Video-MME, the benchmark that anchored mid-2024 comparisons, has been superseded in frontier-model reporting by Video-MMMU, which measures not merely perception but knowledge acquisition from expert-level videos across six disciplines [61,62]. Among the three models compared here, Google’s Gemini 3 Pro pairs the largest context window—one million tokens, allowing entire films, hour-long recordings, or large document sets to be supplied in a single prompt without chunking or subsampling—with a vendor-reported Video-MMMU score of 87.6% [55].

OpenAI’s GPT-5.2 operates with a smaller context window (400k tokens) yet reported 90.5% on Video-MMMU, overtaking Gemini 3 Pro on the accuracy metric while trailing it on raw context capacity [57,63]; on OpenAI’s multi-needle long-context retrieval evaluation (MRCR, eight needles at 128k tokens), practitioner analyses report a mean match ratio of 85% for GPT-5.2 against 77% for Gemini 3 Pro [63]. Anthropic’s Claude Opus 4.5 retains a 200k-token window and is not primarily positioned for long-form video [56,64]. That the accuracy leader and the context-capacity leader diverged within three weeks of each other’s release—GPT-5.2 shipped in December 2025 in direct response to Gemini 3 Pro’s November launch—is a concrete instance of the ranking inversion that this section could previously describe only in the abstract. Deployment characterizations—Gemini 3 Pro for archival search, media asset management, and security analytics requiring maximum single-prompt capacity; GPT-5.2 for accuracy-critical video question answering within moderate context—derive from vendor documentation and practitioner benchmarks rather than controlled studies, and are therefore indicative rather than independently validated [10,63].

5.2. Document and Visual Reasoning: Precision and Logic

For tasks requiring precise extraction and logical reasoning over complex visual documents (e.g., user interfaces, scientific charts, and invoices), the measurement landscape has changed even more than the models: DocVQA and MathVista, which differentiated the mid-2024 generation, are now effectively saturated at the frontier, and evaluation has migrated to harder instruments—MMMUI and MMMUI-Pro for multi-discipline visual reasoning [65], CharXiv for realistic chart understanding [66], and ScreenSpot-Pro for fine-grained interface grounding. On MMMUI-Pro, OpenAI reports 86.5% for GPT-5.2 (90.1% with Python tool use), ahead of Gemini 3 Pro’s 81.0% [55,57,63]; on CharXiv reasoning, GPT-5.2 reports 88.7% with tool use [63]. Anthropic’s Claude Opus 4.5 self-reports 80.7% on MMMUI [64] and was the first model to exceed 80% on SWE-bench Verified [56]; practitioner literature accordingly recommends it for Intelligent Document Processing (IDP), long-document extraction, and software-engineering agents, characterizations that originate in vendor announcements and should be treated as descriptive [56]. Gemini 3 Pro’s 72.7% on

ScreenSpot-Pro, far ahead of its peers, positions it for agents that must ground actions in rendered user interfaces [10].

Two measurement caveats deserve emphasis. First, these figures are predominantly vendor-reported. Second—and new to this generation—third-party re-evaluations of the same models under different harnesses yield materially different scores: Anthropic’s Claude Opus 4.6 system card, for example, reports GPT-5.2 at 79.5–80.4% on MMMU-Pro under its own evaluation harness, roughly six points below OpenAI’s self-reported figure [60]. Cross-vendor benchmark claims are thus sensitive not only to model version and date but to the evaluating harness, which reinforces the multi-benchmark, time-stamped reporting standard adopted throughout this section.

5.3. Audio Intelligence: Beyond Transcription

Historically, audio processing relied on deterministic pipelines: audio was transcribed to text using a speech-to-text model (like Whisper), and the resulting text was then analyzed. This method inherently loses non-lexical acoustic signals, such as sarcasm, emotional hesitation, and background noise. In the current generation, native omni-modal processing is no longer a differentiator but a baseline: Gemini 3 Pro accepts interleaved text, image, audio, and video inputs within its million-token window, allowing it to “hear” acoustic nuances and perform content-aware speaker handling without an external diarization model [55,67], and GPT-5.2 processes audio natively in conversational deployments and through OpenAI’s real-time interfaces [57]. Claude’s multimodal support remains centered on text and vision, with audio typically handled through a transcription stage [56].

Practitioner comparisons continue to attribute the lowest-latency voice interaction to OpenAI’s real-time audio stack, making GPT-5.2 a frequently recommended choice for conversational agents, while Gemini 3 Pro’s long context window is described as advantageous for extended meeting summarization, call-center analysis, and security audio monitoring [10,63]. Whisper remains a widely used, cost-efficient option for pure high-volume transcription, though it lacks the acoustic-reasoning capabilities embedded in natively multimodal models. It should be noted that, because no standardized public benchmark comparable to Video-MMMU or MMMU-Pro yet anchors audio-intelligence evaluation, the claims in this subsection rest more heavily on vendor and practitioner sources than those in Sections 5.1 and 5.2, and should be weighted accordingly.

Table 2 summarizes functionalities and use cases of major frontier models in terms of multimodality. However, when evaluating the relative capabilities of frontier models such as Google Gemini, OpenAI GPT, and Anthropic Claude, the findings must be interpreted with methodological caution. First, competition within the foundation model ecosystem is extraordinarily intense; providers continuously execute rapid, iterative updates that can invert model rankings within weeks. A snapshot of SOTA performance at any single point in time quickly becomes obsolete. Second, the evaluations themselves vary drastically depending on the specific multimodal and reasoning metrics employed. While one architecture may lead on general deep-research benchmarks like GAIA or specialized coding and issue-resolution suites like SWE-bench Multimodal, a different architecture might excel on advanced reasoning evaluations such as Humanity’s Last Exam or the Abstraction and Reasoning Corpus (ARC-AGI). Furthermore, aggregated industry telemetry from independent platforms like Artificial Analysis highlights that factors such as pricing efficiency and tool-use latency introduce completely separate vectors of optimization. Consequently, cross-model analysis requires a holistic, timely, and continuously updated multi-benchmark paradigm rather than reliance on a single, scalar metric [68–70].

Table 2. Comparative Analysis of Frontier Multimodal Models.

Model	Primary Strength	Key Benchmark	Use Case
Google Gemini 3 Pro (November 2025)	Long-context multimodal ingestion (1 M tokens); native audio and video	Video-MMMU: 87.6%; MMMU-Pro: 81.0% (November 2025; vendor-reported [55]). ScreenSpot-Pro: 72.7% (practitioner-compiled [10])	Media archival, long-recording analysis, GUI-grounded agents (vendor/practitioner-reported [10])
OpenAI GPT-5.2 (December 2025)	Visual, chart, and video reasoning accuracy; agentic tool use	Video-MMMU: 90.5%; MMMU-Pro: 86.5% (90.1% with tools) (December 2025; vendor-reported [57,63]; competitor’s harness reports ~79.5–80.4% [60])	Accuracy-critical document/video QA, professional knowledge work (vendor/practitioner-reported [63])
Anthropic Claude Opus 4.5 (November 2025)	Long-document processing; agentic coding workflows	MMMU: 80.7% (November 2025; vendor-reported [64]); SWE-bench Verified: 80.9% (vendor-reported [56])	Intelligent document processing, software-engineering agents (vendor-reported [56])

Note: Scores are time-stamped snapshots for the specific model versions listed; incremental successors (Gemini 3.1 Pro Preview, Claude Opus 4.6, GPT-5.5) had been released by mid-2026 and relative rankings may have changed [58–60]. “Vendor-reported” denotes figures from provider announcements, model cards, or system cards that have not been independently replicated; where an independent re-evaluation under a different harness exists, it is cited alongside. The use-case column reflects vendor documentation and practitioner commentary and is indicative rather than validated.

There is a deeper analytical point that emerges from setting Section 5 against the architectures of Sections 3 and 4: The binding constraint in multimodal data management is increasingly retrieval, not generation. The benchmark gains catalogued above measure what a frontier model can do once the relevant evidence is in its context window; yet recent evaluations show that the same models perform strikingly poorly when used as multimodal retrievers—generative giants remain retrieval weaklings [51]—and high-dimensional nearest-neighbor search itself carries long-standing theoretical caveats about the meaningfulness of distance in high dimensions [71,72]. The practical implication is that improvements on the generation benchmarks of this section do not translate automatically into end-to-end improvements for the retrieval-centric pipelines of Sections 3 and 4; organizations evaluating a model upgrade should benchmark the full retrieve-then-reason loop on their own corpora, not the model in isolation [42].

Readers who wish to go beyond the survey level of this review are directed as follows: For the theoretical foundations of multimodal representation, alignment, and fusion, Baltrušaitis et al. [27] and Liang et al. [28]; for architectural surveys of large multimodal models, including training recipes and efficiency techniques, Yin et al. [30] and Jin et al. [35]; and for the retrieval-centric and agentic extensions on which Sections 4 and 7 build, Zheng et al. [46] and Singh et al. [21].

6. Governance and Operational Challenges

Whereas Section 2 through Section 5 addressed the structural and semantic limitations of unstructured data management, this section turns to the third strand of the review’s thread—the epistemic limitation: the scalability, interpretability, verification, and governance problems introduced by probabilistic Systems of Reasoning. The transition to probabilistic Systems of Reasoning introduces risks that deterministic Systems of Record never faced. The major issue is that scalability in multimodal systems is non-linear. This creates the following challenges:

1. Indexing Costs: Storing billions of 4096-dimensional vectors requires immense RAM. The “Curse of Dimensionality” means that as dimensions increase, distance metrics

(like Euclidean distance) become less meaningful, requiring more expensive computations to find true nearest neighbors [71,73,74].

2. Latency vs. Recall: Approximate Nearest Neighbor (ANN) algorithms (like HNSW) trade accuracy for speed. In high-stakes domains (e.g., legal discovery), missing a relevant document due to index approximation is unacceptable. Systems must balance this by using Re-Ranking: retrieving a larger set of candidates quickly and then using a heavy cross-encoder model to re-rank them accurately [75].

3. Data Gravity: Moving terabytes of video archives from on-premise storage to the cloud for embedding is often cost-prohibitive due to egress fees [76]. This is driving Edge AI, where models (like Gemini Nano) run locally to filter and embed data at the source, sending only the vectors (metadata) to the cloud [77].

6.1. Interpretability, Hallucination, and Independent Modality Verification

The “Black Box” nature of embeddings is a major barrier to adoption in regulated industries. For example, why did the model think this video of a “handshake” was similar to “corruption”? The embedding space does not provide an explanation. Techniques like Concept-Based Explanations are emerging to probe these spaces, but they are in their infancy. In multimodal RAG, models can “hallucinate” details that are not in the retrieved context, especially when the visual input is cluttered or low-resolution. Studies show that hallucination rates in citation generation can be significant. Implementing “Citation Evaluation” loops where a second model verifies that the generated claim is actually supported by the retrieved pixel/text data is becoming a standard architectural pattern [78,79].

To mitigate the risk of stochastic errors and “hallucinations” in complex AI systems, researchers increasingly advocate for Independent Modality Verification (IMV) as a framework for high-assurance decision-making. Under an IMV architecture, disparate data streams, or “modalities”, operate as isolated sensory inputs to validate a single proposition, ensuring that the system’s confidence is grounded in multi-source consensus rather than a single, potentially corrupted pipeline [41,80,81].

The core strength of IMV lies in its structural decoupling. In traditional multimodal fusion, features are often integrated early in the process (“Early Fusion”), which allows a failure in one input to propagate through the entire neural network. In contrast, IMV utilizes a Late Fusion or Decision-Level Fusion approach. This ensures that the manipulation or “noise” affecting one modality does not automatically bias the other, providing a robust defense against adversarial attacks and sensor failure [82].

As specified in the three-component model, a typical IMV workflow consists of [83,84]:

1. Modality A (e.g., Facial Recognition): Captures and processes visual biometric data.
2. Modality B (e.g., Voice Recognition): Captures and processes auditory/linguistic data.
3. Verification Logic: A symbolic or rule-based controller that evaluates the outputs.

For the system to authorize a decision, both modalities must independently exceed a predefined confidence threshold and reach a congruent conclusion. This creates a “redundant safety” mechanism similar to those found in aerospace and medical diagnostics.

Despite its security benefits, IMV introduces significant computational overhead [85]. Running parallel inference pipelines increases latency and hardware requirements, which may be prohibitive for edge computing devices with limited power budgets. Furthermore, the implementation of IMV is governed by the principle of Adaptive Security. For certain low-risk applications, such as unlocking a mobile device for non-sensitive tasks, the “convenience-to-security ratio” favors single-factor authentication. High-rigor IMV is typically reserved for “High-Stakes AI” environments, such as autonomous driving

or financial transactions, where the cost of a false positive is catastrophic [86,87]. The workflow of IMV is illustrated in Figure 2.

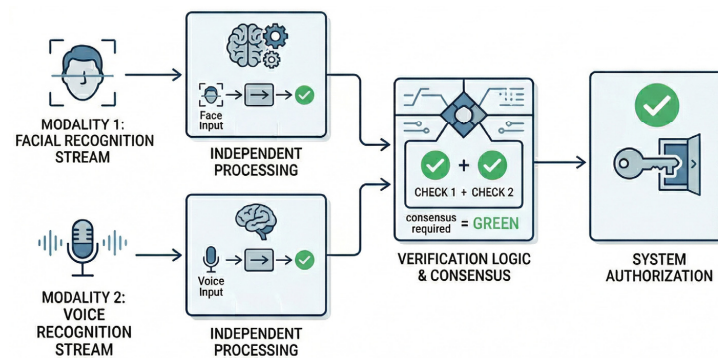


Figure 2. Independent Modality Verification (IMV) Framework.

6.2. Governance: PII, Bias, and Unlearning

A critical vulnerability in current data management is the industry misconception that high-dimensional vector embeddings operate as one-way cryptographic hashes, rendering the data mathematically anonymized. Historically, system architects treated vector databases as low-risk storage. However, recent empirical research into embedding inversion attacks comprehensively refutes this assumption [74,88]. Adversaries can successfully reverse-engineer dense text embeddings to reconstruct the exact original input. Utilizing a multi-step controlled generation process that iteratively corrects hypothetical text against a target embedding space, researchers were able to recover the majority of the original token text sequences. Most critically for regulatory compliance, this inversion process successfully recovered highly sensitive personal attributes—including exact first and last names—from datasets of clinical medical notes. Further studies confirm that without the application of differential privacy techniques (such as injecting calculated statistical noise into the vector generation process), both textual and visual embeddings remain vulnerable to extraction. Consequently, vector embeddings must not be treated as anonymized meta-data; they require the same level of encryption, access control, and regulatory protection as the raw plaintext or original imagery they represent [89,90]. To rectify the situation, organizations must implement PII Masking (using tools like Microsoft Presidio) before the data are embedded [22]. Additionally, Private Information Retrieval (PIR) protocols are being explored to allow querying vector databases without revealing the query itself.

Moreover, models like CLIP inherit biases from their internet-scale training data. They might associate images of doctors primarily with men. To remediate the problem, modern algorithms allow engineers to identify specific dimensions in the vector space that correlate with protected attributes (gender, race) and “prune” or re-weight them, reducing bias without retraining the entire model [78].

The realization that vectors contain reconstructable personal data necessitates strict lifecycle governance for embeddings, particularly regarding consent management and the GDPR’s “Right to be Forgotten” (Article 17) [91]. When a user revokes consent or a data retention policy expires, simply deleting the source document from a cloud bucket is legally insufficient. A compliant architecture must execute synchronized deletion across three distinct layers: the raw file storage, the semantic vector database, and the model’s parametric memory. To address the latter, researchers are advancing machine unlearning protocols. Frameworks such as SISA (Sharded, Isolated, Sliced, and Aggregated) training allow system administrators to isolate and purge specific user data segments without the computationally prohibitive requirement of retraining the entire foundational model from scratch. By implementing strict Role-Based Access Control (RBAC) at the vector-retrieval

layer—ensuring an agent cannot retrieve an embedding if the querying user lacks clearance for the source document—organizations can maintain compliance throughout the entire lifecycle of the multimodal data [92].

7. Future Horizons

While Section 3 through Section 6 described technologies mature enough for production deployment, this section surveys three emerging directions—world models, neuro-symbolic integration, and personal intelligence ecosystems—that extend the same trajectory from perception toward reasoning.

7.1. World Models and Any-to-Any Training

The field is evolving rapidly beyond static perception. Consequently, relying solely on traditional proficiencies—such as relational database administration or unimodal programming—is increasingly inadequate for managing highly heterogeneous data ecosystems. While current models predict the next token, future “World Models” will predict the state of the world. Specifically, a World model is a predictive AI architecture that learns an internal representation of the physical world’s dynamics—including causality, physics, and object permanence—enabling it to simulate and anticipate outcomes in an abstract representation space rather than merely predicting the next token or pixel.

To actualize the vision of world models, Yann LeCun founded Advanced Machine Intelligence (AMI) Labs, focusing on the Joint-Embedding Predictive Architecture (JEPA) [93]. Unlike LLMs that predict pixels or words, JEPA predicts the state of the world in an abstract representation space. These models are trained primarily on video and sensory data rather than just text, allowing them to learn “common sense” and basic physics. Doing so can enable them to ignore “noise” (like the exact texture of a leaf) and focus on “signal” (like the trajectory of a falling object). They will understand physics, causality, and object permanence, allowing them to simulate outcomes (“What happens if I drop this glass?”) rather than just describing pixels [94,95]. This approach is crucial for robotics and autonomous systems [96]. In addition, newer approaches are moving away from models trained on fixed pairs (image-text) to models trained on unified streams of data. “Any-to-Any” models will be able to translate between different modalities directly (e.g., humming a melody to generate a violin score) without passing through a text bottleneck, leading to more fluid and human-like interaction.

7.2. Neuro-Symbolic Integration and GraphRAG

Another trend in data management for unstructured data is the reintegration of symbolic AI. While Neural Networks (Deep Learning) are excellent at perception (identifying a cat in a video), they are poor at strict logic and rule adherence. Neuro-Symbolic AI combines the best of connectionism and symbolism [97], an approach that is manifested in GraphRAG [98]. Recent work has also explored neuro-symbolic AI in advanced air mobility [99]. GraphRAG is a retrieval-augmented generation approach that combines vector-based semantic search with knowledge-graph traversal, using a language model to extract entities and relationships from the corpus and organize them into a structured graph, thereby enabling explicit, traceable reasoning over multi-hop relationships that pure vector similarity would miss. Instead of just retrieving vector chunks, the GraphRAG approach uses an LLM to extract nodes (entities) and relationships from the data to build a Knowledge Graph. Retrieval then involves traversing this graph. Because the relationships are explicit, the reasoning behind an answer is traceable and thus explainability is greatly enhanced [73,100]. For example, in a financial investigation, vector search might find documents mentioning “Company A” and “Company B.” A Knowledge Graph can reveal

that “Director X” sits on the board of both, a relationship that might be three hops away and invisible to vector similarity. GraphRAG combines the “fuzzy” search of vectors with the “structured” pathfinding of graphs.

7.3. Personal Intelligence (PI) and Resource Integration

The three ecosystems discussed in this subsection—Microsoft Copilot, Google’s Personal Intelligence, and Apple Intelligence—are presented as illustrative cases of one candidate direction in unstructured data management: the integration of structured and unstructured resources (emails, cloud documents, meeting transcripts, multimedia libraries) into a unified semantic index at the ecosystem level. That three major providers have independently converged on broadly similar architectures is suggestive, but it should not be read as evidence of a settled architectural paradigm; these systems are recent, proprietary, and largely unevaluated by independent research, and their long-term viability remains an open question. They are examined here because they concretize, at consumer and enterprise scale, several of the architectural patterns analyzed in Sections 3 and 4.

Because these ecosystems are proprietary and no peer-reviewed evaluations of them yet exist, the descriptions in this subsection are based on vendor documentation and announcements [67,101–104]; they should be read as accounts of vendor-stated capabilities rather than independently validated evidence, consistent with the source-handling policy described in Section 1.2.

In enterprise environments, the integration of multimodal resources is largely driven by tenant-grounded assistants that map corporate data silos. For example, Microsoft’s Copilot operates as a multimodal AI platform that spans Windows, Microsoft 365 applications, and Bing. It utilizes Microsoft Graph to index organizational content—such as SharePoint documents, Teams meeting transcripts, and Outlook inbox context—allowing the system to retrieve and synthesize data via Retrieval-Augmented Generation (RAG). To handle external data, this ecosystem employs “synced connectors” to ingest and semantically index external content into Microsoft Graph, or “federated connectors” to retrieve real-time data using the Model Context Protocol (MCP) without moving the underlying data. For example, Microsoft’s documentation describes Copilot ingesting an unstructured meeting transcript and cross-referencing it with a structured CRM database to draft negotiation briefs [102]. Furthermore, Microsoft states that the MAI model family will allow organizations to adapt models to local workflows via ‘Frontier Tuning’ via localized Reinforcement Learning Environments (RLEs). This mechanism allows the underlying foundational models to ingest and adapt directly to specific user workflows, local institutional data, and cross-modal resources securely within the enterprise’s isolated tenant environment [103].

Personal Intelligence (PI), often referred to as the “PT” (Personal Technology), is Google’s ecosystem-level AI framework that integrates a user’s data across interconnected services (Gmail, Google Drive, Google Photos, YouTube) into a unified, context-aware reasoning system capable of retrieving and synthesizing information across these sources [67]. By granting PI access to an ecosystem that includes Gmail, Google Drive, Google Photos, and YouTube, the system builds a “knowledge graph.” This allows the AI to reason across silos, connecting a flight confirmation in Gmail to a presentation slide in Drive and a captured photo in Google Photos. Unlike previous iterations of AI that treated each query as an isolated event, PI functions as a continuous memory-based assistant that “remembers” the user’s professional context, research interests, and digital history [101].

We use the term Research Data Intelligence (RDI) to denote a conceptual extrapolation—proposed here, not an existing product—of ecosystem-level intelligence from personal accounts to research data management. The scenarios that follow, and the demonstrations in Section 9, illustrate what such a capability looks like within the current, consumer-scale implemen-

tation of PI; whether it generalizes to institutional settings is an empirical question. The transition toward Research Data Intelligence (RDI) represents a substantial methodological advancement for researchers managing unstructured data. Traditionally, managing complex datasets required specialized technical skills in relational databases, such as SQL, or non-relational structures, such as NoSQL. PI lowers this barrier by allowing researchers to simply upload any type of multimodal data directly to Google Drive, from Word documents, PDFs, PowerPoint slides, and spreadsheets, to PNG/JPEG images, and MP3/MP4 files. In many instances, even the act of “uploading” is automated. For example, photos taken during fieldwork are synchronized to Google Photos, and a researcher’s YouTube viewing history, such as watching documentaries on global policy, is automatically indexed as a searchable context [67]. This architecture facilitates a continuously updated dataset, wherein data ingestion and indexing are managed by native ecosystem synchronization rather than manual metadata entry. The distinguishing technical feature of PI is its native multimodal processing. While conventional systems rely on human-generated labels, tags, or text-based transcripts to “understand” media, Gemini utilizes a unified transformer architecture that “sees” image pixels, “listens” to audio waves, and “watches” video frames directly [55]. By converting visual and auditory signals into the same token-based representational format used for text, the model achieves a much higher level of nuance. It can identify the architectural style of a building in a photo or the frustrated tone in an interview subject’s voice without needing a text-based description first [67].

Contrasting the cloud-heavy approaches of Microsoft and Google, alternative ecosystems prioritize on-device processing to manage multimodal personal resources. Apple Intelligence, powered by Apple Foundation Models, embeds personal context understanding directly into the operating system across iOS, iPadOS, and macOS. Rather than centralizing data in a cloud data warehouse, this architecture utilizes a local semantic index (Spotlight) built from user entities across Messages, Mail, and Photos. A defining feature of this ecosystem is “onscreen awareness,” which allows the system’s agent (Siri AI) to answer queries related to the unstructured visual and textual content currently displayed on a user’s screen. When complex multimodal reasoning requires more compute than edge devices can provide, the architecture leverages “Private Cloud Compute” to process requests without persistently storing or sharing the user’s data [104].

While the capability to synthesize emails, cloud drives, and multimedia is actively being deployed, discussions regarding future enterprise capabilities must remain grounded in current operational realities. The deployment of integrated ecosystem intelligence is not an automatic process; it introduces profound challenges in data governance, compliance, and access control. For instance, utilizing RAG to surface historical contract clauses or meeting transcripts within a platform like Microsoft Copilot requires strict role-based permissions and conditional access to prevent data leakage or the exposure of sensitive Personally Identifiable Information (PII). Furthermore, because foundational models are probabilistic, the risk of hallucination remains a critical barrier. Industry best practices currently dictate that while multimodal AI can draft briefs or synthesize historical data, human verification and sign-off remain mandatory for critical enterprise outputs. Consequently, while the vision of a fully autonomous, cross-ecosystem reasoning agent is compelling, the immediate future of enterprise multimodal data management will heavily depend on structured governance, strict access controls, and “human-in-the-loop” verification architectures.

8. Taxonomy

To synthesize the diverse advancements discussed in this review, we propose a structured taxonomy of multimodal data management (Table 3). A mature System of Reasoning is not defined merely by adopting a frontier foundation model, but by aligning all seven di-

mensions of this framework. For instance, an advanced architecture must pair omni-modal capabilities with Zero-ETL ingestion, secure it with SISA-style governance, and execute queries via Agentic RAG [72,105].

Table 3. A Structured Taxonomy of Multimodal Data Management Systems.

Architectural Dimension	Legacy/Deterministic Systems	Transitional/Unimodal AI	Advanced Multimodal Systems (Systems of Reasoning)
Modality	Structured (Tabular/Numerical)	Siloed Unstructured (Text, Image, Audio)	Omni-modal/Cross-modal (Any-to-Any)
Ingestion Strategy	Manual ETL Pipelines	Separated Storage & Compute (Cloud buckets)	Zero-ETL/Active Pointers (e.g., BigQuery ObjectRef)
Representation Method	Relational Schemas (SQL)	Feature Engineering/Specific Classifiers	High-Dimensional Semantic Embeddings & Knowledge Graphs
Retrieval Architecture	Keyword/ID Lookup	Standard RAG (Static, one-shot retrieval)	Agentic RAG & GraphRAG (Dynamic routing, re-ranking)
Reasoning Layer	Static Dashboards/Rules	Single-model Inference	Multi-agent Orchestration & Independent Modality Verification (IMV)
Governance Mechanism	Role-Based Access Control	Static PII Filtering	Private Information Retrieval (PIR) & SISA Machine Unlearning
Deployment Context	On-Premise Relational DB	Fragmented Cloud Data Swamps	Multimodal Lakehouse & Native Ecosystems (e.g., Google PI/RDI)

By categorizing systems along these axes, researchers and enterprise architects can systematically identify bottlenecks in their data pipelines—such as utilizing advanced embeddings (Representation) but relying on static, one-shot lookups (Retrieval)—thereby transitioning from disjointed AI tools to integrated Research Data Intelligence (RDI).

9. Discussion: Use Cases

The four cases below are drawn deliberately from a consumer-grade implementation—Google’s Personal Intelligence (PI)—because it is, at the time of writing, the most complete publicly accessible instance of the always-on, multimodal, retrieval-centric pattern this review has described. Their evidentiary status is addressed directly in Section 9.5: they are demonstrations of feasibility, not validated research methodology.

9.1. Textual Data and Ethics Research

The first author of this article is a professor of Data Science and AI Ethics. In this case, PI facilitates the management of dense philosophical and technical literature. The professor can store thousands of pages regarding the major schools of moral philosophy, such as Eudaimonist, Deontological, and Utilitarian ethics, within a dedicated Google Drive folder. By querying PI, the researcher can instantly synthesize how these frameworks apply to modern “black box” algorithmic transparency. Furthermore, these data can be connected to NotebookLM, an AI-native research tool that allows for even deeper grounded analysis. In this workflow, PI acts as the retrieval engine that fetches relevant passages across multiple files, which NotebookLM then uses to generate citations and structured summaries for an academic paper or grant proposal [101].

9.2. Audio Data and Qualitative Fieldwork

In the realm of qualitative research, PI substantially streamlines the analysis of interviews, journals, and other forms of open-ended data. Traditionally, researchers had to

manually transcribe audio or rely on third-party services like Otter.ai, Sonix, or Descript. These transcripts were then exported to specialized qualitative software like MAXQDA or IBM Modeler for coding and textual analytics. In the RDI framework, researchers can simply upload MP3 recordings to Google Drive. Then PI can analyze the audio natively, identifying themes, emotional cues, and speaker intent without the intermediate step of generating a separate text file. This is particularly useful for linguistic research or psychological studies where the way something is said (prosody and pitch) is as important as the words themselves [67].

9.3. Visual Data and Ethnography

For sociologists and artists, PI turns the smartphone into an ethnographic data-collection instrument. A researcher studying the architectural styles of Honolulu's Chinatown can simply walk through the neighborhood taking photos. Because the device is linked to Google Photos, the images are automatically uploaded and indexed by PI before the researcher even returns to their desk. PI can then be asked to "Find all photos of buildings with Richardsonian Romanesque lava rock" or "Summarize the evolution of signage in the historical district." This automates the "coding" process that previously took weeks of manual labor, allowing the researcher to focus on higher-level cultural synthesis [101]. Appendix A shows an example of how Google's PI is utilized to retrieve and analyze visual data.

9.4. Video Data and Technical Analysis

Video data management becomes considerably more efficient when integrated with YouTube and PI. If a researcher wants to perform a comparative study of hardware, like comparing GPU, TPU, and LPU architectures, they can search for and view relevant technical videos on YouTube. Because the viewing history is recorded by PI, the AI can "watch" these movies on behalf of the user to provide a technical summary of the performance benchmarks mentioned across different channels. Additionally, a researcher could upload recorded technical demonstrations to Drive; PI can then identify specific moments where a system failure occurred or summarize the visual steps of a complex software installation, bridging the gap between instructional video and technical documentation [67]. Appendix B shows an example of using PI to retrieve and analyze video data from YouTube. However, the outputs reproduced in the appendices are provided solely for illustrative purposes and were not used as empirical evidence to assess the system's accuracy, completeness, or reliability. Nonetheless, the factual claims, external links, and classifications contained in these outputs were independently verified.

9.5. From Demonstration to Method: Governance, Reproducibility, and Bias Control

The demonstrations above establish that a Research Data Intelligence (RDI) workflow is technically feasible on commodity infrastructure. They do not establish that it is methodologically adequate, and three gaps separate the former from the latter.

The first gap is governance. Research data ingested into a consumer platform is governed by the platform's terms of service rather than by an institutional data-management plan. Interview recordings, journals, and field photographs routinely contain identifiable participants, so ethics-board protocols (e.g., IRB), informed-consent language, and the applicable regulatory regimes (GDPR, HIPAA, FERPA) must be satisfied before upload—a condition a consumer service cannot certify on the researcher's behalf. The risks documented in Section 6 apply with full force in this setting: because embeddings are invertible, participant data that have been indexed cannot be treated as redacted [74,88,89], and honoring a participant's withdrawal engages the machine-unlearning problem rather than simple file deletion [91,105]. In practice, an RDI workflow involving human-subject data

belongs on enterprise or institutional deployments that offer contractual data-processing terms, retention control, region pinning, and audit logging—the consumer implementation demonstrated here is appropriate only for non-sensitive, researcher-owned material.

The second gap is reproducibility. Foundation models are stochastic, and consumer services update silently, so the same query over the same corpus may return different syntheses weeks apart, with no record of which model version produced which output. A reportable RDI workflow therefore requires, at minimum: logging the model identifier, version, and query date for every analysis; archiving prompts and raw outputs alongside the derived findings; preferring API access with pinned model versions over consumer front-ends wherever results will be published; and re-running key analyses to estimate output stability. Absent these controls, PI-style findings remain demonstrations rather than evidence.

The third gap is bias control. AI-generated codes, themes, and summaries inherit both the documented biases of large multimodal models [78] and their propensity to hallucinate [41,80]. The established validity apparatus of qualitative research—codebook auditing, negative-case analysis, and inter-rater comparison between machine-generated and human codes—therefore remains necessary rather than obsolete, and earlier work on the compatibility of text mining with grounded theory and content analysis provides a template for such triangulation [5,14]. On this view, AI-assisted coding should be reported as a first-pass procedure whose outputs are verified by human analysts, with the machine–human disagreement rate disclosed as a standard quality indicator.

9.6. *Systems of Reasoning*

The demonstrations in this section illustrate, at the scale of a single researcher’s account, the pattern this review has termed a System of Reasoning: a platform that synthesizes multimodal data streams rather than passively storing them. Whether unified intelligence architectures such as Microsoft’s Copilot, Google’s PI, and Apple Intelligence can carry this pattern into the enterprise data stack remains to be demonstrated. If they do, the transition will not be a “plug-and-play” upgrade: it would require re-architecting the data platform around ObjectRef-style storage, vector indices, and agentic workflows, embedded in a governance framework that accounts for the probabilistic and opaque character of neural systems. The 80% of organizational data that remains unstructured represents the potential payoff of such a transition, not a guaranteed outcome of it.

10. Conclusions and Future Directions

This review began from a structural fact of the contemporary data landscape: An estimated 80% of enterprise and research data now exists in unstructured, multimodal forms that traditional schema-driven architectures were never designed to manage. The central problem, we have argued, is no longer storage but meaning. NoSQL systems resolved the structural limitation of the relational paradigm—they can hold a million videos or clinical notes—but they cannot read, hear, or watch what they hold. The key issue for modern data management is therefore semantic: how to build systems that understand, integrate, and reason across heterogeneous modalities, and how to govern the probabilistic interpretations such systems produce.

The review developed four core ideas in response. First, the multimodal embedding serves as the unifying substrate that projects text, images, audio, and video into a shared semantic space, transforming unstructured data from a passive liability into a queryable asset. Second, operationalizing this substrate requires a three-stage pipeline—ingestion (exemplified by Zero-ETL patterns such as ObjectRef), representation (modality-aware chunking and cross-modal embedding), and vector storage (the multimodal lakehouse).

Third, retrieval-augmented generation (RAG), and increasingly its agentic variants, is the mechanism that connects stored representations to model reasoning; yet this does not displace structured databases, which remain the auditable, ACID-compliant anchor of trustworthy AI pipelines. Fourth, the resulting shift is best characterized as a movement from Systems of Record, which report what happened, to Systems of Reasoning, which synthesize evidence across modalities to explain why.

This review leads to three conclusions as follows: (1) Multimodal AI and RAG have crossed the threshold of production readiness, whereas world models and neuro-symbolic architectures remain promising but exploratory; adoption decisions should reflect this maturity gradient. (2) Governance is a precondition rather than an afterthought: because embeddings are invertible and models hallucinate, mature deployments must pair semantic capability with PII masking, machine unlearning, independent modality verification, and access control at the retrieval layer. (3) A mature System of Reasoning is defined not by adopting a frontier model but by alignment across all seven dimensions of the taxonomy proposed in Section 8—modality, ingestion, representation, retrieval, reasoning, governance, and deployment—which practitioners can use diagnostically to locate bottlenecks in their own architectures.

Looking forward, PI's ability to bridge traditionally siloed platforms such as Gmail, Drive, Photos, and YouTube could, in principle, be extended to an enterprise context to form an integrated knowledge graph—the Research Data Intelligence framework proposed in Section 7.3. Realizing it would require context to be preserved and enriched across multimodal interactions rather than fragmented across systems, and would inherit all of the governance preconditions discussed in Section 6. For long-form video archives, Gemini 3 Pro's one-million-token context window makes it a practical candidate for sustained analysis of lengthy content without complex segmentation or sliding-window techniques, consistent with the Video-MMMU results reviewed in Section 5.1 [55,61]. For legal and technical document processing, Claude Opus 4.5's document-oriented positioning and agentic tooling (Section 5.2) suggest suitability for tasks demanding high extraction precision [56,64], while GPT-5.2's chart- and document-reasoning results position it for accuracy-critical analytical workloads [57,63]. All three recommendations are tied to the specific model versions and evaluation dates discussed in Section 5 and should be re-validated against current benchmarks at the time of adoption.

From an infrastructure perspective, transitioning toward a multimodal lakehouse architecture is advisable. Platforms such as BigQuery with Vertex AI or vector-native systems like LanceDB enable tight coupling between raw data and vector representations, aligning with the PI model of continuous and automated data synchronization. This approach reduces the dependency on traditional ETL pipelines, which often introduce latency, redundancy, and operational complexity. At the same time, the challenge of data gravity, where large datasets are costly to move, can be mitigated by adopting edge AI strategies. By relocating computation closer to where data is generated, organizations can process and filter information locally, transmitting only compact representations such as embeddings to centralized systems.

Equally important is the need for robust governance and regulatory compliance. Techniques such as SISA-style data sharding and systematic PII masking provide a scalable foundation for privacy-preserving architectures. Organizations can design systems that enforce strict data boundaries while still enabling powerful cross-modal intelligence by drawing inspiration from the implicit “air gaps” within PI. For example, such separation may exist between public-facing YouTube content and private Studio data. This balance between accessibility and control is critical for meeting GDPR and other regulatory requirements.

This discussion is by no means exhaustive, particularly given the rapid pace of innovation in AI. New models, architectures, and best practices continue to emerge on a near-weekly basis. As such, practitioners and researchers alike should actively monitor developments, experiment with evolving techniques, and remain adaptable in order to effectively navigate the shifting landscape of multimodal AI.

Author Contributions: Conceptualization: C.H.Y.; Validation: C.H.Y., N.M. and Z.W.; Resources: C.H.Y., N.M. and Z.W.; Writing—Original Draft Preparation: C.H.Y.; Writing—Review & Editing: C.H.Y., N.M. and Z.W.; Supervision: C.H.Y.; Project Administration: C.H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the preparation of this manuscript/study, the authors used ChatGPT 5.5 and Google Gemini 3.1 for the purposes of brainstorming, information gathering, proofreading, and use case illustration. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RAG	Retrieval-Augmented Generation
LLM	Large Language Model
SQL	Structured Query Language

Appendix A. Example of Using Google’s PI to Retrieve and Analyze Visual Data

Prompt: “Please access my account on photos.google.com. Please select only the photos taken in Chinatown, Honolulu, view those photos and summarize their architectural style.”

Gemini’s answer:

Based on your images, the architectural style of Honolulu’s Chinatown can be summarized as a complex, multicultural hybrid of 19th-century Western commercial forms and traditional Asian cultural motifs. Here are the key themes I identified:

1. 19th-Century Western Commercial Architecture

Many of the buildings you photographed (particularly the two-story brick and stone structures) are in the Italianate or Renaissance Revival styles.

- **Key Features:** You captured high, narrow windows with arched lintels, decorative cornices along the rooflines, and sturdy masonry facades—some using the local “blue stone” (volcanic basalt).
- **Significance:** These reflect the rebuilding of Chinatown after the devastating fires of 1886 and 1900, showing the influence of Western developers during the territorial era.

2. Traditional Asian Vernacular & Symbolic Motifs

Interspersed with the Western frames are clear markers of the neighborhood’s cultural identity.

- **Roofs and Ornamentation:** I see photos of the classic upturned eaves (swallowtail roofs) and green glazed tiles that characterize traditional Chinese architecture, often found on temples or association buildings.
- **Color Palette:** There is a recurring use of “luck and prosperity” colors—deep reds for pillars and gold for decorative inscriptions and signage.
- **The Dragon Pillars:** Your photos of the ornate dragon-wrapped pillars at major intersections or market entrances are iconic examples of modern traditionalism.

3. The “Mixed-Use” Market Vernacular

A significant portion of your photos focuses on the open-air market entrances and street-level storefronts.

- **Function-Driven Design:** These feature large folding metal gates and recessed entryways, designed to facilitate the high-traffic flow of the produce and fish markets.
- **Aesthetic:** It’s a “lived-in” style where the architecture serves as a backdrop for vibrant, hand-painted signage and the display of goods, creating a dense, textured urban environment.

4. Historical “Time-Capsule” Elements

You also caught some of the faded signage and ghost murals on the sides of older buildings. These provide a visual timeline of the neighborhood’s economic history, showing how storefronts have transitioned from early 20th-century saloons to modern-day boutiques and galleries (see Figure A1).

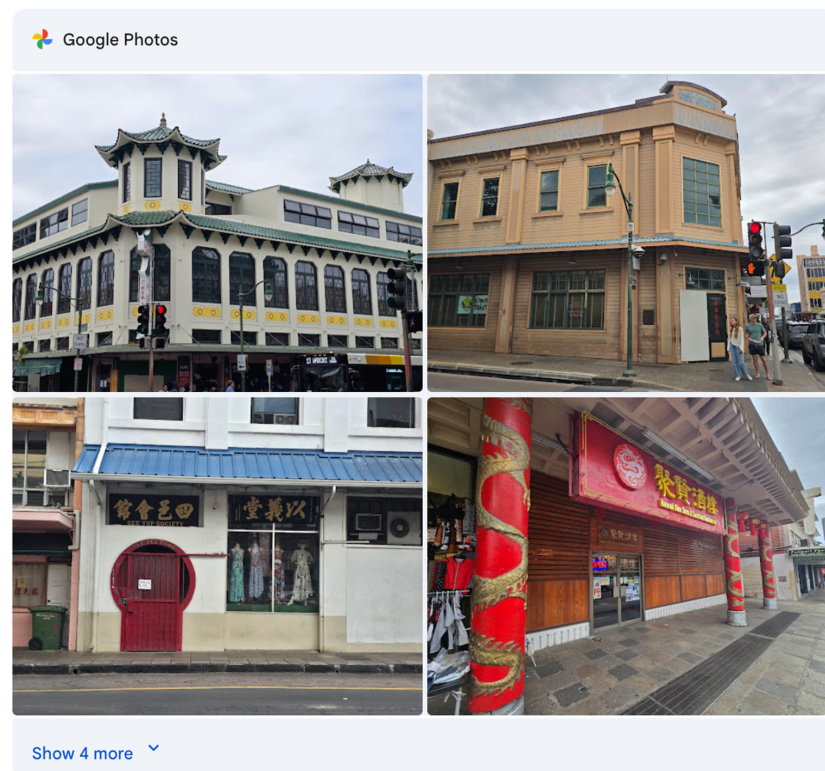


Figure A1. Visual data of architecture in Chinatown, Honolulu. In the lower-left image, the characters on the left sign read “See Yup Society”; “See Yup” (also romanized as Sze Yup or Siyi, meaning “Four Counties”) is a historical region in the southern part of Guangdong Province, China. The characters on the right sign mean “Hall of Righteousness,” the name of a sworn-brotherhood lodge. The characters on the sign in the lower-right image mean “Gathering of Worthies Wine House,” or more loosely, “A Restaurant Where Wise and Talented People Gather.” The remaining Chinese signage in the figure is incidental storefront text and is not referenced in the analysis.

Appendix B. Example of Using PI to Retrieve and Analyze Video Data from YouTube

Prompt: “Access the YouTube videos about comparing Nvidia GPU and Google TPU. I want to know how many videos favor TPU, how many favor GPU, and how many are neutral? Please show one example of each category (i.e., the link to the video).”

Gemini’s answer:

Based on a representative sample of 10 prominent videos from late 2024 and 2025, the breakdown of stances is as follows:

- Favoring Google TPU (3 videos): These videos often frame the TPU as an “Nvidia-killer,” focusing on “ruthless efficiency” and the breakthrough in performance-per-dollar for massive LLM training.
- Favoring Nvidia GPU (2 videos): These prioritize the “CUDA moat” and architectural flexibility, arguing that GPUs remain a better choice for the broader AI market (beyond just language models).
- Neutral/Comparative (5 videos): These provide balanced technical deep dives, framing the choice as a strategic “Implementation Detail” based on whether a project needs R&D flexibility or production-scale efficiency.

Representative Examples and Analysis

1. Favoring Google TPU: “Google’s New TPU Quietly Ends the GPU Era?” This video presents the TPU v5p as a direct and superior challenger to Nvidia’s flagship hardware. Google’s TPUs discard general-purpose computing entirely to optimize strictly for matrix multiplication using a hardware structure called a systolic array. It views Google’s vertical integration (chip + software + cloud) as a game-changer that challenges Nvidia’s market dominance.

- Link: https://www.youtube.com/watch?v=iseM_Sb-ruo (accessed on 10 May 2026)

2. Favoring Nvidia GPU: “IT’S OVER! I Can’t Stay Quiet on GOOGLE vs. NVIDIA Any Longer” While the title is sensational, the analysis concludes that Nvidia’s “general-purpose” architecture makes it the safer and more versatile bet for 90% of the industry. It argues that while TPUs win on “token generation” for Google and Meta, GPUs are the only viable hardware for robotics, protein folding, and self-driving cars. It frames the TPU as a niche alternative that only aims to capture 10% of the market.

- Link: <http://www.youtube.com/watch?v=ZZ2nWg1QhR4> (accessed on 25 March 2026)

3. Neutral/Balanced: “TPU vs. GPU: The Ultimate Guide to the AI Hardware War” This video provides a highly technical comparison, describing GPUs as “Swiss Army Knives” for messy R&D and TPUs as “Industrial Apple Corers” for ruthless production efficiency. It posits that the “war” is not zero-sum, but rather a diversification of the supply chain. It concludes that with modern compilers like JAX and XLA, the “hardware choice” is becoming less of a technical prison and more of a boardroom decision based on unit economics.

- Link: <http://www.youtube.com/watch?v=THruaU8V6gg> (accessed on 25 March 2026)

Appendix C. Example of Using PI to Retrieve and Analyze Video Data on Google Drive

Prompt: “In the folder “Unit 5” of “6500” in my Google drive, please summarize the video about how quantum technology would affect security of cloud computing.”

Gemini’s answer:

The video in your “Unit 5” folder, titled “The Quantum Dawn: When Encryption Meets Its Greatest Challenge,” presented by Alex Yu, explores how the transition from classical to quantum computing creates a fundamental shift in cloud security.

The Quantum Challenge to Encryption

The core message is that current digital security—protecting everything from financial transactions to confidential cloud files—relies on mathematical problems that are currently “unbreakable” by classical supercomputers. Quantum technology changes this landscape through several key mechanisms:

- **Breaking Mathematical Foundations:** Current schemes like RSA and ECC rely on the difficulty of factoring large prime numbers. Shor’s algorithm, designed for quantum processors, can solve these problems exponentially faster, effectively dismantling the foundation of modern encryption.
- **Weakening Symmetric Encryption:** Grover’s algorithm can accelerate brute-force attacks, halving the effective key strength of symmetric encryption like AES.
- **“Harvest-Now, Decrypt-Later”:** A significant emerging risk is that adversaries may collect and store encrypted cloud traffic today with the intent of decrypting it once quantum tools become sufficiently mature.

Impact on Cloud Infrastructure

Cloud computing’s entire trust framework—including TLS/SSL connections, user authentication, and data-at-rest encryption—is built on these vulnerable cryptographic methods. The video warns that once quantum computing reaches full maturity, these frameworks could essentially collapse, making current data protection methods obsolete.

The Race Toward Quantum-Resistant Security

To mitigate these risks, the video discusses two primary defensive paths:

- **Post-Quantum Cryptography (PQC):** Developing new encryption standards (like CRYSTALS-Kyber and CRYSTALS-Dilithium) that use “lattice-based puzzles”—complex geometric structures believed to be resistant to quantum attacks.
- **Quantum Key Distribution (QKD):** A defense strategy that uses the laws of physics rather than mathematics. Because observing a quantum particle changes its state, any attempt to eavesdrop on a message is immediately detectable.

The video concludes that while the “Quantum Dawn” poses a threat to digital privacy, it also offers a “Quantum Renaissance” for breakthroughs in areas like medicine and materials science, provided humanity can adapt its security frameworks fast enough.

References

1. Yu, C.H. *Philosophical Foundations of Quantitative Research Methodology*; University Press of America: Lanham, MD, USA, 2006.
2. Yu, C.H. *Philosophical Foundations of Classical Statistic, Bayesian Statistic, and Data Science/Machine Learning*. In *Philosophical Principles of Quantitative Research*; Liebe, U., Ed.; Edward Elgar Publishing: Northampton, MA, USA, 2026; in press.
3. Husserl, E. *The Crisis of European Sciences and Transcendental Phenomenology*; Carr, D., Translator; Northwestern University Press: Evanston, IL, USA, 1936/1970.
4. Schutz, A. *The Phenomenology of the Social World*; Walsh, G.; Lehnert, F., Translators; Northwestern University Press: Evanston, IL, USA, 1967.
5. Denzin, N.K.; Lincoln, Y.S. *The SAGE Handbook of Qualitative Research*, 5th ed.; SAGE Publications: Thousand Oaks, CA, USA, 2018.
6. Yu, C.H. Are positive trait attributions for the deceased caused by fear of supernatural punishments?: A triangulated study by content analysis and text mining. *J. Psychol. Christ.* **2015**, *34*, 3–18.
7. Jain, A.K.; Ross, A.; Nandakumar, K. *Introduction to Biometrics*; Springer: Berlin/Heidelberg, Germany, 2011.
8. Jain, A.K.; Flynn, P.J. *Handbook of Biometrics*; Springer: Berlin/Heidelberg, Germany, 2019.
9. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [[CrossRef](#)] [[PubMed](#)]
10. Zeeb, N. Google Gemini 3 Benchmarks (Explained). Vellum AI. Available online: <https://www.vellum.ai/blog/google-gemini-3-benchmarks> (accessed on 10 August 2026).

11. Talukdar, W.; Hu, J.; Mascarenhas, K.; Zhang, L. Unleashing the Multimodal Power of Amazon Bedrock Data Automation. AWS Machine Learning Blog. Available online: <https://aws.amazon.com/blogs/machine-learning/unleashing-the-multimodal-power-of-amazon-bedrock-data-automation-to-transform-unstructured-data-into-actionable-insights/> (accessed on 20 March 2025).
12. Yu, C.H. *Data Mining and Exploration: From Traditional Statistics to Modern Data Science*; CRC Press: Boca Raton, FL, USA, 2022.
13. Yu, C.H.; DiGangi, S.; Jannasch-Pennell, A. Using text mining for improving student experience management. In *Cases on Innovations in Educational Marketing: Transnational and Technological Strategies*; Tripathi, P., Mukerji, S., Eds.; IGI Global: Hershey, PA, USA, 2011; pp. 196–213.
14. Yu, C.H.; Jannasch-Pennell, A.; DiGangi, S. Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *Qual. Rep.* **2011**, *16*, 730–744.
15. Yu, C.H.; Jannasch-Pennell, A.; DiGangi, S. Enhancement of student experience management in higher education by sentiment analysis and text mining. *Int. J. Technol. Educ. Mark.* **2018**, *8*, 16–33. [\[CrossRef\]](#)
16. Han, J.; E, H.; Le, G.; Du, J. Survey on NoSQL database. In Proceedings of the 2011 6th International Conference on Pervasive Computing and Applications, Port Elizabeth, South Africa, 25–29 October 2011; pp. 363–366. [\[CrossRef\]](#)
17. Microsoft. Multimodal Search in Azure AI Search. Microsoft Learn. 2025. Available online: <https://learn.microsoft.com/en-us/azure/search/multimodal-search-overview> (accessed on 25 March 2026).
18. Poccia, D. Amazon Nova Multimodal Embeddings Now Available in Amazon Bedrock. AWS News Blog. Available online: <https://aws.amazon.com/blogs/aws/amazon-nova-multimodal-embeddings-now-available-in-amazon-bedrock/> (accessed on 28 October 2025).
19. Su, J.; Soni, G. Introducing BigQuery ObjectRef: Supercharge Your Multimodal Data and AI Processing. Google Cloud Blog. Available online: <https://cloud.google.com/blog/products/data-analytics/new-objectref-data-type-brings-unstructured-data-into-bigquery> (accessed on 25 June 2025).
20. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
21. Singh, A.; Chen, J.; Zhang, Y.; Wang, W.Y. Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv* **2025**, arXiv:2501.09136.
22. Chen, J.; Bian, V. Safe RAG with HydroX AI and Zilliz: PII Masking for Responsible GenAI. Zilliz Blog. 2024. Available online: <https://zilliz.com/blog/safe-rag-with-hydrox-ai-and-zilliz-pii-masking-for-responsible-genai> (accessed on 25 March 2026).
23. Baethge, C.; Goldbeck-Wood, S.; Mertens, S. SANRA—A Scale for the Quality Assessment of Narrative Review Articles. *Res. Integr. Peer Rev.* **2019**, *4*, 5. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Ferrari, R. Writing Narrative Style Literature Reviews. *Med. Writ.* **2015**, *24*, 230–235. [\[CrossRef\]](#)
25. Greenhalgh, T.; Thorne, S.; Malterud, K. Time to Challenge the Spurious Hierarchy of Systematic over Narrative Reviews? *Eur. J. Clin. Invest.* **2018**, *48*, e12931. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Sukhera, J. Narrative Reviews: Flexible, Rigorous, and Practical. *J. Grad. Med. Educ.* **2022**, *14*, 414–417. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Liang, P.P.; Zadeh, A.; Morency, L.-P. Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *ACM Comput. Surv.* **2024**, *56*, 264. [\[CrossRef\]](#)
29. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A Survey on Multimodal Large Language Models. *Natl. Sci. Rev.* **2024**, *11*, nwae403. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Gandomi, A.; Haider, M. Beyond the Hype: Big Data Concepts, Methods, and Analytics. *Int. J. Inf. Manag.* **2015**, *35*, 137–144. [\[CrossRef\]](#)
32. Armbrust, M.; Ghodsi, A.; Xin, R.; Zaharia, M. Lakehouse: A New Generation of Open Platforms That Unify Data Warehousing and Advanced Analytics. In Proceedings of the 11th Conference on Innovative Data Systems Research (CIDR 2021), Online, 11–15 January 2021. Available online: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf (accessed on 10 May 2026).
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
34. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML 2021), Virtual, 18–24 July 2021; Volume 139, pp. 8748–8763.
35. Jin, Y.; Li, J.; Gu, T.; Liu, Y.; Zhao, B.; Lai, J.; Gao, Z.; Wang, Y.; Wang, C.; Tan, X.; et al. Efficient multimodal large language models: A survey. *Vis. Intell.* **2025**, *3*, 27. [\[CrossRef\]](#)

36. LanceDB. *Designed for Multimodal: Built for Scale*; LanceDB: San Francisco, CA, USA, 2025. Available online: <https://lancedb.com/> (accessed on 25 March 2026).
37. Levandoski, J.; Casto, G.; Deng, M.; Desai, R.; Edara, P.; Hottelier, T.; Hormati, A.; Johnson, A.; Johnson, J.; Kurzyniec, D.; et al. BigLake: BigQuery's evolution toward a multi-cloud lakehouse. In *Proceedings of the 2024 International Conference on Management of Data (SIGMOD'24)*, Santiago, Chile, 9–15 June 2024. Available online: <https://research.google/pubs/biglake-bigquerys-evolution-toward-a-multi-cloud-lakehouse/> (accessed on 25 March 2026).
38. Xie, J.; Zhou, X.; Cheng, L. Edge computing for real-time decision making. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 598–607.
39. Shashanka, B.R.; Charan, M.; Banu, F.S. Chunking strategies for multimodal AI systems. *arXiv* **2025**, arXiv:2512.00185.
40. Gao, Y.; Ji, Z.; Zhao, G. A Multimodal retrieval-augmented generation system for intelligent question answering. In *Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Prague, Czech Republic, 9–12 October 2022*; IEEE: New York City, NY, USA, 2022; pp. 104–110.
41. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
42. Mortaheb, M.; Khojastepour, M.A.A.; Chakradhar, S.T.; Ulukus, S. RAG-Check: Evaluating Multimodal Retrieval Augmented Generation Performance. *arXiv* **2025**, arXiv:2501.03995.
43. Wang, Y.; Zhang, Z. Visual-RAG: Benchmarking Text-to-Image Retrieval Augmented Generation for Visual Knowledge Intensive Queries. *arXiv* **2025**, arXiv:2502.16636v2.
44. Kim, J.; Tao, R.; Sharma, S.; Wang, J.; Sun, K.; Lin, Z.; Moon, S.; Mathias, L.; Kumar, A.; Ji, H.; et al. Pixel-grounded retrieval for knowledgeable large multimodal models. *arXiv* **2026**, arXiv:2601.19060v1.
45. Samanta, H. Grounded Multimodal Retrieval-Augmented Drafting of Radiology Impressions Using Case-Based Similarity Search. *arXiv* **2026**, arXiv:2603.17765v1.
46. Zheng, X.; Weng, Z.; Lyu, Y.; Jiang, L.; Xue, H.; Ren, B.; Paudel, D.; Sebe, N.; Gool, L.V.; Hu, X. Retrieval augmented generation and understanding in vision: A survey and new outlook. *arXiv* **2025**, arXiv:2503.18016.
47. Xia, P.; Zhu, K.; Li, H.; Wang, T.; Shi, W.; Wang, S.; Zhang, L.; Zou, J.; Yao, H. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. *arXiv* **2024**, arXiv:2410.13085v2.
48. Liu, B.; Ning, Y.; He, S.; Guo, F.; Jia, S.; Zhu, L. A Multimodal Retrieval-Augmented Generation System for Intelligent Question Answering. In *Cognitive Computing—ICCC 2025; Lecture Notes in Computer Science*; Yang, Y., Huang, M., Pan, X., Zhang, J., Chen, J., Zhang, L.J., Eds.; Springer: Cham, Switzerland, 2026; Volume 16156. [CrossRef]
49. Raymond, B. Level Up Your GenAI Apps: What's Next for RAG. Unstructured.io Blog. Available online: <https://unstructured.io/blog/level-up-your-genai-apps-what-s-next-for-rag> (accessed on 4 June 2025).
50. Liu, N.F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; Liang, P. Lost in the middle: How language models use long contexts. *arXiv* **2023**, arXiv:2307.03172.
51. Feng, H.; Sheng, Z.; Qiang, M.; Zhang, W. Generative giants, retrieval weaklings: Why do multimodal large language models fail at multimodal retrieval? *arXiv* **2025**, arXiv:2512.19115.
52. Kumar, A. Hands-on Agentic RAG. Medium. 2025. Available online: <https://abvijaykumar.medium.com/hands-on-agentic-rag-1-2-cdf375ad7e7a> (accessed on 25 March 2025).
53. Microsoft. Microsoft SQL Server. 2025. Available online: <https://www.microsoft.com/en-us/sql-server> (accessed on 25 March 2026).
54. Gray, J.; Reuter, A. *Transaction Processing: Concepts and Techniques*; Morgan Kaufmann: Burlington, MA, USA, 1993.
55. Google. Gemini 3: Introducing the Latest Gemini AI Model from Google. Available online: <https://blog.google/products-and-platforms/products/gemini/gemini-3/> (accessed on 7 August 2026).
56. Anthropic. Introducing Claude Opus 4.5. Available online: <https://www.anthropic.com/news/claude-opus-4-5> (accessed on 8 August 2026).
57. OpenAI. Introducing GPT-5.2. Available online: <https://openai.com/index/introducing-gpt-5-2/> (accessed on 10 August 2026).
58. Bristot, D. Gemini 3.1 Pro Preview: What the .1 Actually Means. WhatLLM. Available online: <https://whatllm.org/blog/gemini-3-1-pro-preview> (accessed on 10 August 2026).
59. OpenAI. Introducing GPT-5.5. Available online: <https://openai.com/index/introducing-gpt-5-5/> (accessed on 10 August 2026).
60. Anthropic. *Claude Opus 4.6 System Card*; Anthropic: San Francisco, CA, USA, 2026. Available online: <https://www-cdn.anthropic.com/0dd865075ad3132672ee0ab40b05a53f14cf5288.pdf> (accessed on 10 August 2026).
61. Hu, K.; Wu, P.; Pu, F.; Xiao, W.; Zhang, Y.; Yue, X.; Li, B.; Liu, Z. Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. *arXiv* **2025**, arXiv:2501.13826.
62. Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-Modal LLMs in Video Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)*, Nashville, TN, USA, 11–15 June 2025; IEEE: New York, NY, USA, 2025; pp. 24108–24118.

63. SotaTek. What Is GPT-5.2? Comparison Between GPT-5.2 vs. Gemini 3 Pro. Available online: <https://sotatek.com.au/blogs/gpt-5-2/> (accessed on 10 August 2026).
64. Anthropic. *Claude Opus 4.5 System Card*; Anthropic: San Francisco, CA, USA, 2025. Available online: <https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf> (accessed on 10 August 2026).
65. Yue, X.; Zheng, T.; Ni, Y.; Wang, Y.; Zhang, K.; Tong, S.; Sun, Y.; Yu, B.; Zhang, G.; Sun, H.; et al. MMMU-Pro: A More Robust Multi-Discipline Multimodal Understanding Benchmark. *arXiv* **2024**, arXiv:2409.02813.
66. Wang, Z.; Xia, M.; He, L.; Chen, H.; Liu, Y.; Zhu, R.; Liang, K.; Wu, X.; Liu, H.; Malladi, S.; et al. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. *arXiv* **2024**, arXiv:2406.18521.
67. Google Cloud. Audio Understanding with Gemini. Google Cloud Vertex AI Documentation. Available online: <https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/audio-understanding> (accessed on 25 March 2025).
68. SWE-Bench. SWE-Bench Multimodal. Available online: <https://www.swebench.com/multimodal.html> (accessed on 7 June 2026).
69. Center for AI Safety & Scale AI. Humanity’s Last Exam. Available online: <https://agi.safe.ai/> (accessed on 7 June 2026).
70. Artificial Analysis. Independent Analysis of AI. Available online: <https://artificialanalysis.ai/> (accessed on 7 June 2026).
71. Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When Is “Nearest Neighbor” Meaningful? In Proceedings of the 7th International Conference on Database Theory (ICDT 1999), Jerusalem, Israel, 10–12 January 1999; pp. 217–235.
72. Malkov, Y.A.; Yashunin, D.A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 824–836. [CrossRef] [PubMed]
73. Dong, D. The Limitations of Vector-Based Retrieval in RAG. 2025. Available online: <https://writer.com/blog/vector-based-retrieval-limitations-rag/> (accessed on 25 March 2025).
74. Morris, J.X.; Kuleshov, V.; Shmatikov, V.; Rush, A.M. Text Embeddings Reveal (Almost) as Much as Text. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), Singapore, 6–10 December 2023; pp. 12448–12460.
75. Gong, W.G. Reranker optimization via geodesic distances on k-NN manifolds. *arXiv* **2026**, arXiv:2602.15860.
76. Barry, C. What Is Data Gravity and Why Does It Matter? Barracuda. Available online: <https://blog.barracuda.com/2026/01/28/what-is-data-gravity-and-why-does-it-matter-> (accessed on 28 January 2026).
77. IBM, Inc. What Is Edge AI? 2026. Available online: <https://www.ibm.com/think/topics/edge-ai> (accessed on 10 May 2026).
78. Adewumi, T.; Alkhaled, L.; Gurung, N.; van Boven, G.; Pagliai, I. Fairness and Bias in Large Multimodal Models: A Survey. *arXiv* **2024**, arXiv:2406.19097.
79. Lakera Team. LLM Hallucinations in 2026: How to Understand and Tackle AI’s Most Persistent Quirk. *Lakera.ai*. 2026. Available online: <https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models> (accessed on 10 May 2026).
80. Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; Shou, M.Z. Hallucination of multimodal large language models: A survey. *arXiv* **2024**, arXiv:2404.18930.
81. Iwanowski, M.; Gahbler, M. Multiple large AI models’ consensus for object detection—A survey. *Appl. Sci.* **2025**, *15*, 12961. [CrossRef]
82. Shao, Z.; Dou, W.; Pan, Y. Dual-level Deep Evidential Fusion: Integrating multimodal information for enhanced reliable decision-making in deep learning. *Inf. Fusion* **2023**, *103*, 102113. [CrossRef]
83. Alharbi, B.; Alshanbari, H.S. Face-voice based multimodal biometric authentication system via FaceNet and GMM. *PeerJ. Comput. Sci.* **2023**, *9*, e1468. [CrossRef] [PubMed]
84. Bengio, S.; Marcel, C.; Marcel, S.; Mariéthoz, J. Confidence measures for multimodal identity verification. *IDIAP Res. Rep.* **2002**, *3*, 267–276. [CrossRef]
85. Li, P.; Wang, X.; Huang, K.; Huang, Y.; Li, S.; Iqbal, M. Multi-model running latency optimization in an edge computing paradigm. *Sensors* **2022**, *22*, 6097. [CrossRef] [PubMed]
86. Szczuko, P.; Harasimiuk, A.; Czyżewski, A. Evaluation of decision fusion methods for multimodal biometrics in the banking application. *Sensors* **2022**, *18*, 2356. [CrossRef] [PubMed]
87. Mahmood, A.; Szabolcsi, R. A systematic review on risk management and enhancing reliability in autonomous vehicles. *Machines* **2025**, *13*, 646. [CrossRef]
88. Tonic.ai. Sensitive Data in Text Embeddings Is Recoverable. Tonic.ai Blog. Available online: <https://www.tonic.ai/blog/sensitive-data-in-text-embeddings-is-recoverable> (accessed on 16 November 2025).
89. Liu, T.; Yao, H.; Wu, T.; Qin, Z.; Lin, F.; Ren, K.; Chen, C. Mitigating Privacy Risks in LLM Embeddings from Embedding Inversion. *arXiv* **2024**, arXiv:2411.05034v1.
90. Chen, Y.; Lent, H.; Bjerva, J. Text Embedding Inversion Security for Multilingual Language Models. *arXiv* **2024**, arXiv:2401.12192.
91. European Data Protection Supervisor. Machine Unlearning. TechSonar. 2025. Available online: https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/machine-unlearning_en (accessed on 25 March 2026).

92. Parihar, D. Your AI Remembers Too Much: How Differential Privacy Can Fix the Vector Database Privacy Crisis. Medium. Available online: <https://dev523.medium.com/your-ai-remembers-too-much-how-differential-privacy-can-fix-the-vector-database-privacy-crisis-96199ef14ed8> (accessed on 7 June 2026).
93. O'Brien, C. Yann LeCun's AMI Labs Launches with \$1.03 Billion to Build AI that Understands the Real World. The French Tech Journal 2026. Available online: <https://www.frenchtechjournal.com> (accessed on 25 March 2026).
94. LeCun, Y. A Path Towards Autonomous Machine Intelligence. OpenReview. 2022. Available online: <https://openreview.net/pdf?id=BZ5a1r-kVsf> (accessed on 25 March 2026).
95. Maes, L.; Le Lidec, Q.; Scieur, D.; LeCun, Y.; Balestriero, R. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv* **2026**, arXiv:2603.19312.
96. Meta. V-JEPA: The Next Step Toward Advanced Machine Intelligence. Meta AI Blog. 2024. Available online: <https://ai.meta.com/blog/v-jepa-yann-lecun-ai-model-video-joint-embedding-predictive-architecture/> (accessed on 25 March 2026).
97. Hitzler, P.; Eberhart, A.; Ebrahimi, M.; Sarker, M.K.; Zhou, L. Neuro-symbolic approaches in artificial intelligence. *Natl. Sci. Rev.* **2022**, *9*, nwac035. [CrossRef] [PubMed]
98. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv* **2024**, arXiv:2404.16130.
99. Acharya, K.; Sharifi, I.; Lad, M.; Sun, L.; Song, H. Integrating neurosymbolic AI in advanced air mobility: A comprehensive survey. *arXiv* **2025**. [CrossRef]
100. Hunger, M. What Is GraphRAG? Neo4j. Available online: <https://neo4j.com/blog/genai/what-is-graphrag/> (accessed on 24 March 2026).
101. Google. Personal Intelligence: Help That's Made for You. Available online: <https://gemini.google/overview/personal-intelligence> (accessed on 10 August 2026).
102. Microsoft. Announcing Microsoft Copilot, Your Everyday AI Companion. Microsoft News Center. Available online: <https://news.microsoft.com/apac/2023/09/21/announcing-microsoft-copilot-your-everyday-ai-companion/> (accessed on 7 June 2026).
103. Suleyman, M. Building a Hill-Climbing Machine: Launching Seven New MAI Models. Microsoft AI Blog. Available online: <https://microsoft.ai/news/building-a-hillclimbing-machine-launching-seven-new-mai-models/> (accessed on 6 June 2026).
104. Apple Developer. Apple Intelligence. Apple Developer Documentation. 2026. Available online: <https://developer.apple.com/apple-intelligence/> (accessed on 6 June 2026).
105. Bourtole, L.; Chandrasekaran, V.; Choquette-Choo, C.A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; Papernot, N. Machine Unlearning. In Proceedings of the 2021 IEEE Symposium on Security and Privacy (S&P), San Francisco, CA, USA, 24–27 May 2021; pp. 141–159.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.