Journal of International DBA Studies (JIDS)
Vol. 2 No. 001 (2026) https://eaj.ebujournals.lu/index.php/JIDS/index

Multimodal Generative AI Agents for Biomedical Document Classification: Architecture, Ethical Boundaries, and Human-in-the-Loop Governance

by

Arun Kumar *

Abstract

The rapid increase in biomedical research publications has made it difficult for researchers, clinicians, and policymakers to efficiently review and interpret scientific information. Traditional manual review methods and rule-based automation tools are no longer sufficient to manage the growing volume, complexity, and multimodal nature of modern biomedical literature, where important insights need to be presented through both written text and visual elements such as figures and images. To address this challenge, this study proposes and evaluates a multimodal generative AI agent for biomedical document classification and image captioning, which combines an instruction-tuned language model with a vision encoder to process abstract text and related visual content together. The agent operates within a controlled framework that includes human oversight to ensure responsible and ethical use. To study its effect a mixed-method approach was used, including quantitative performance evaluation and qualitative expert review. The model was tested on open-access biomedical papers from arXiv across four subject areas. Results indicate that the multimodal approach performs better than text-only systems in classification accuracy and contextual understanding. However, the findings also show that human supervision remains important in order to reduce risks related to bias and incorrect outputs. This study therefore offers practical and theoretical guidance for developing ethical and reliable AI systems in biomedical research settings.

Keywords: Multimodal Generative AI, Biomedical Document Classification, Large Vision-Language Models, AI Agents, Human-in-the-loop(HITL), Bounded Autonomy, AI Governance, Ethical Artificial Intelligence.

*Arun Kumar is a technology leader with 30+ years of global experience who has led large-scale, portfolio-level transformation initiatives across industries including supercomputing, telecom, automotive, retail, and high-technology. He brings a rare combination of deep technical expertise and strategic business acumen, with advanced qualifications including an MTech in Data Science & Machine Learning (PES University), an MBA in Business Analytics (BITS Pilani), and executive education in Technology Leadership & Innovation from MIT. He is currently pursuing a Doctorate in Business Administration (Emerging Technologies – Generative AI) at Golden Gate University

Introduction

The amount of biomedical research published each year has risen sharply. Databases such as PubMed now contain more than 35 million indexed records, with new studies added on a daily basis. This steady expansion makes it difficult for researchers, clinicians, policymakers, and healthcare institutions to keep up with current findings and identify reliable evidence in a timely way. In addition to the increase in volume, the structure of biomedical articles has become increasingly more complex. Many studies now combine detailed written explanations with figures, medical images, charts, and experimental diagrams. Reviewing such material manually requires significant time and effort, and the risk of oversight or inconsistency increases as the number of publications grows.

As a result, there is a clear need for intelligent systems that can help organize and interpret large collections of biomedical information while preserving accuracy, transparency, and ethical responsibility.

Traditional methods for analyzing biomedical literature have largely relied on single-modality approaches. Some systems focus only on text using natural language processing techniques to classify abstracts or extract keywords. Others process images separately through automated visual analysis tools. When text and images are examined independently, important relationships between written explanations and visual evidence may be overlooked. This gap is particularly significant in biomedical research, where figures, charts, and microscopic images often carry essential scientific meaning.

Recent multimodal systems have attempted to combine text and images within a unified framework. However, many of these systems operate as performance-driven models. Typically, they emphasize classification accuracy or caption quality without giving sufficient attention to interpretability, accountability, or practical governance. In biomedical contexts, where incorrect interpretation can impact research direction or clinical support, these limitations are not trivial. To address this risk, the approach presented in this study moves beyond simple multimodal integration. Instead, it frames the system as a task-oriented AI agent that functions under defined operational boundaries. Human oversight is embedded into the workflow to review uncertain or high-risk outputs. By combining multimodal reasoning with structured supervision and governance, the proposed design offers a more controlled and responsible alternative to existing multimodal systems.

Specifically, to correct these limitations, this paper designs and analyzes the architecture of a multimodal generative AI agent through the application of two language models, namely vision-language tuned and instruction-tuned, with the help of the integration of two pipelines, which operate jointly to perform biomedical image captioning and abstract-level text categorization. To test the accuracy, robustness, interpretability, and management effect, a mixed-method study design is taken to the extent that quantitative performance evaluation and qualitative expertise analysis are adopted. The human-in-the-loop checks, ethical AI issues, and regulatory practices that could be adopted to introduce biomedical AI implementation are given a special mention, which could be used to offer AI implementation a gentle introduction.

As such this paper can be said to make contributions to the literature in three ways. The first is a redefinition of the multimodal generative AI system is that moves it beyond just a limited autonomous analytical system toward a weak autonomous agent. Second, it gives objective data about the benefits of multimodal integration as opposed to a unimodal process of comprehending biological texts. Third, it provides a viable framework of equality between the intangible ideas of AI agents and formal administrative and ethical mandates. All these investments singly contribute to the knowledge-intensive domains of biology when they are combined in the effectual use of AI agents and effective interpretation of a phenomenon.

Literature Review

Automated Biomedical Document Classification

The idea of automated biomedical document image categorization is not a new phenomenon of research activity; in fact, it has more than 20 years history. With the development of the sphere of scientific life and medicine is also becoming exponentially bigger. At the same time, earlier models that were adapted on the basis of systems rule-based and classical machine learning methods like topic modelling, naive Bayes classifiers, and support vector machines were problematic; these techniques limited extrapolation and application to sub-fields of biomedicine because of their reliance manually generated features and domain-specific heuristics with limited automation.

From these earlier origins, the art of biomedical text categorization has improved tremendously with the emergence of deep learning. Transformer-based learning contextual representations have been proven to be more efficient in comparison to convolutional neural networks (CNNs) (Devlin et al., 2019), recurrent neural networks (RNNs), and earlier models based on data learning. As part of its improvement, domain-adapted language model, BioBERT and PubMedBERT, was linked to the substantial increase in classification because of the ability to use long biomedical corpora. However, the majority of text-based still remain limited due to the impossibility of taking into account extraneous data in tables, figures, and experimental images, which are necessary, in most cases. to understand biomedical works.

Biomedical Image Analysis and Captioning

The analysis of biomedical images has though been greatly boosted quickly alongside text analysis using the deep learning methods used in radiology, pathology, microscopy, and biological imaging. The models based on CNN have been effective in image classification, image segmentation, and image detection. However, whereas biological picture comprehension requires semantic interpretation, other contexts rely instead on the detection of patterns.

It has thus been suggested that the image captioning models may be utilized as a model of visual-textual linkage in that they are capable of generating the natural-language description of the visual contents. The encoders of the original images were CNNs, and the decoders, which utilized an RNN architecture or transformer

architecture, could be used in biomedical captioning. When applied to specialized biological images in this way, these models have been prone to poor contextual grounding, including linguistic errors, even though they made the interpretations more comprehensible. At worst, they can conceal written description that can compromise scientific context and the objective nature of experiments.

Multimodal Learning for Biomedical Documents

The new study has taken into account the multimodal learning techniques in which both the text and visuals are modeled simultaneously and attempt to overcome the points of limitation of the unimodal systems noted above; multimodal architecture that is better applied in the tasks of document categorization, information extraction, and visual question answering, portray certain possibilities of more infused semantic links when applied in clinical accounts, pathological images, and academic writings in biomedical literature.

Nonetheless, it is on the basis of these developments that there are still a number of issues that remain. To begin with, there are multimodal systems in which their modalities are asymmetrical; inference is dominated by one of them. Second, the evaluation occurs without proper assessment of interpretability, robustness, and operation, by emphasizing only measures of accuracy. Third, unlike reusability, multimodal systems have been structured into workflow-integrated autonomous systems. These drawbacks complicate their use in real biomedical applications.

Generative AI and Vision–Language Models

Generative AI (and especially machine vision large language models) have also been added to multimodal systems. Despite the ability of vision-language models to provide smooth descriptions with the help of visual interaction, it is possible to perform few-shot and zero-shot categorizations using instruction-tuned LLAM. This is because such techniques are efficient to adjust in the new areas imminent in a brief period as well as their reliance on large labeled datasets (FLODA, 2024; Penny-Wise and Pound-Foolish in AI-Generated Image Detection, 2024)

However, there are other risks that go hand-in-hand with the generative models, such as the lack of determinism, bias proliferation, and hallucinations. Such risks are intensified in biomedical settings since they tend to undermine clinical decision-making and future policy-making procedures. For this reason, the use of purely autonomous generative models remains controversial; this has led to an appeal in the need for regulations that would be able to address innovation and responsibility.

Agents and Bounded Autonomy

AI agents can provide a common platform in the incorporation of such multimodal creative competencies into working systems. Generally speaking, AI agents may be characterized as having the capacity to see the surrounding space; make decisions relying on the information that they have at a particular moment; and take actions

according to the set objectives. Agents possess a situation-adapted and directed behavior.

In recent literature, it is important to differentiate between fully autonomous agents and limited autonomous agents who act within a specific realm of authority and are supervised, instead of those that are placed within the boundaries of stated authority. Lack of autonomy has been seen as a major issue for the application of morally oriented AI within closely regulated sectors such as health and finance. Nonetheless, the number of empirical studies assessing AI agents on biomedical knowledge management is not widespread with even fewer studies that consider multimodal generative models as a component of AI controlled agents.

Human-in-the-Loop and Ethical Governance

The current risks linked with AI autonomy need to be reduced by using human-in-the-loop (HITL) techniques. HITL paradigms have the capability of controlling, balancing, and discarding AI outputs since the system is operated by human specialists needed to ensure that such data are not in violation of organizational, legislative, and ethical principles. Research conducted in the past has pointed out greater levels of confidence, accountability, and adoption of the HITL systems in high stakes settings (e.g., Gartner 2024; McKinsey 2025 AI in Healthcare Reports).

Still, rather than adopting HITL as an important design construct, a significant proportion of the existing literature considers it as an afterthought. The recommendation on HITL checkpoints, which can be incorporated into the operations of the AI agents systematically, is not specifically provided concerning the multimodal generative systems.

Research Gap and Motivation

Recent studies have though reported strong progress in generative artificial intelligence, multimodal learning, image analysis, and automated text classification. These advances have improved the technical ability of systems to process complex biomedical data.

Despite this progress, important gaps remain. Many existing multimodal models focus predominately on performance outcomes such as accuracy, while paying limited attention to interpretability, governance, and operational control. In addition, most systems are designed as technical tools rather than as structured agents that function within defined boundaries. This limitation becomes significant in biomedical settings, where decisions must be transparent and subject to human review.

There is also limited research that brings together multimodal generative models, constrained autonomy, and human oversight within a single, organized framework. As a result, practical guidance for deploying such systems responsibly into biomedical knowledge management remains insufficient. To address these gaps, this study seeks to

present a multimodal AI agent design that combines text–image reasoning with structured supervision and governance mechanisms.

Research Methodology

Research Design and Approach

To evaluate the effectiveness, reliability, and sensitivity of a multimodal full-sized AI generative agent when classifying biomedical documents and picture captioning, it is introduced as an explanatory evaluation research design, which is a mixed-method approach.

The document classification process is modeled as a probabilistic decision task. For a given document x, the predicted class $\hat{y}$is determined by selecting the class with the highest posterior probability:

$$\hat{y} = \arg \max_{c \in \{1,\dots,C\}} p(c \mid x)$$

$$p(c \mid x) = \text{softmax}(z)_c = \frac{e^{z_c}}{\sum_{k=1}^{C} e^{z_k}}$$

Equation 1 — Document Classification (Softmax)

Explanation

- $\hat{y}$→ The predicted class label for input x.
- $\arg \max$ → Means "choose the class c" that gives the highest probability.
- $p(c \mid x)$→ The probability that input $x$belongs to class c.
- z_c→ The raw score (logit) output by the model for class c.
- Cdenotes the total number of categories.
- **Softmax function** → Converts raw scores (z_c) into normalized probabilities across all classes C

In short: Equation 1 defines the mathematical rule - AI agent uses to classify documents — it picks the category with the highest Softmax probability. The model assigns the document to the class with the highest predicted probability.

Quantitative evaluation focuses on classification accuracy, scalability, and caption quality. Qualitative assessment examines interpretability, trust, ethical considerations, and managerial usefulness through structured expert feedback.

Construction of Corpus and Data

The empirical analysis of the article is implemented on the premises of open-access publications on arXiv, especially, by quantitative biology (q-bio). Four common

categories were used to select with an opinion to show evidence of the biological diversity and complexity of visuals:

• q-bio.TO – Tissues and Organs

• q-bio.GN – Genomics

• q-bio.CB – Cell Behavior

q-bio.PE -Populations and Evolution.

Sample Size:

A total of 1,200 documents were collected, with 300 documents per category. This balanced sampling strategy was adopted to reduce class imbalance and to ensure equal representation across subject areas.

Data Split:

The dataset was divided into:

- 70% training set (840 documents)
- 15% validation set (180 documents)
- 15% test set (180 documents)

Stratified sampling was applied to maintain proportional class distribution across splits.

Document Structure:

Each document includes:

- A structured abstract (text input)
- At least one associated figure

The corpus contains 1–3 figures per document, depending on availability.

Image Characteristics:

Images include:

- Microscopy images
- Biological pathway diagrams
- Experimental schematics
- Data visualization plots

Image formats primarily include PNG and JPEG. All images were resized to a standardized resolution of 224×224 pixels for consistency with the vision encoder input requirements. Pixel normalization was applied to scale intensity values between 0 and 1.

Text Preprocessing

Abstract text was:

- Lowercased
- Tokenized using the model's tokenizer
- Cleaned to remove metadata and non-content artifacts
- Truncated or padded to a fixed maximum sequence length

Data Quality and Filtering

To ensure dataset quality:

- Documents without figures were excluded
- Corrupted or unreadable image files were removed
- Duplicated entries were filtered out
- Only English-language abstracts were retained

This filtering process resulted in a balanced and clean dataset suitable for multimodal training and evaluation.

Every document of the corpus is having a written abstract and a number of one or more graphics related to it (e.g., pictures of a microscope, biological schematics, or experimental drawings). It is implicated in practiced forms in the biomedical articles in the real world where the meaning of information is disseminated between modalities and is not simply written (Utilizing Generative AI for Patient Behavioral Assessment, 2024). The sample is selected in such a manner that the issue of the sample being biased on the topic is also prevented and we have a well represented sample in the categories.

Artificial Intelligence Agent Pipeline and System Architecture

The proposed solution is designed as a task-oriented multimodal AI agent rather than a collection of separate analytical tools. The system integrates text understanding, image interpretation, and controlled decision-making within a structured workflow. The aim is to evaluate how such an agent can support biomedical knowledge management under realistic operational constraints.

Model Configuration

For textual reasoning and document classification, the system employs an instruction-tuned large language model. The selected model contains approximately 8–70 billion parameters, depending on deployment configuration. In this study, the model is used in a zero-shot setting, meaning it is not retrained on the biomedical dataset. Instead, task-specific prompts guide the classification process. This decision reflects real-world organizational conditions, where retraining large models may not be feasible due to cost, governance, or computational limitations.

For visual processing, a transformer-based vision–language model is used. The image encoder component contains approximately 300 million to 1 billion parameters, depending on the selected configuration. The model converts visual inputs into embeddings that can be aligned with textual representations. Similar to the language model, the vision component is applied in its pre-trained form without domain-specific fine-tuning.

Using pre-trained models allows the study to assess how general-purpose multimodal systems perform in biomedical contexts without additional customization.

System Pipeline

The agent operates through four structured stages:

1. Document Ingestion and Parsing

Biomedical publications are processed to extract abstracts and associated figures. Text is cleaned and tokenized, and images are resized and normalized to ensure consistent input formats.

2. Multimodal Reasoning Engine

Text embeddings produced by the language model are combined with visual embeddings generated by the image encoder. This shared representation enables the system to relate written descriptions to visual evidence within the same document.

3. Decision Generation

The system assigns each document to the category with the highest predicted probability. A confidence score accompanies each classification. These scores help determine whether additional human validation is necessary.

4. Human-in-the-Loop Review

When confidence levels fall below a defined threshold or when ambiguity is detected, outputs are reviewed by domain experts. This mechanism ensures that the system operates under constrained autonomy rather than full automation. It also supports responsible use in regulated biomedical environments.

Evaluation Metrics and Quantitative Analysis

Classification performance is evaluated using standard performance metrics:

- Accuracy
- Precision
- Recall
- Macro-averaged F1-score

Macro-averaging is applied to give equal importance to each document category.

The learning objective is defined using cross-entropy loss:

$$\mathcal{L}_{cls} = -\sum_{c=1}^{C} y_c \log\left(p(c \mid x)\right) \qquad (2)$$

where y_c represents the true class label and $p(c \mid x)$ is the predicted probability.

$\boldsymbol{C}$ **:** This represents the total number of possible document categories (for example, four q-bio classes).

$\boldsymbol{y_c}$ **:**This is the true label for class c.

It uses a one-hot encoding format:

- If the document belongs to class c, then $y_c = 1$
- For all other classes, $y_c = 0$

So only the correct class contributes to the loss.

$\boldsymbol{p(c \mid x)}$ **:** This is the predicted probability that document $x$belongs to class c.

It is computed using the softmax function from the model's output.

$\boldsymbol{\log\left(p(c \mid x)\right)}$ **:** The logarithm is applied to the predicted probability.
If the model assigns high probability to the correct class, the log value is close to 0, and the loss is small.
If the model assigns low probability to the correct class, the log value becomes large (negative), and the loss increases.

The Negative Sign (−) : The negative sign ensures that the loss value is positive. Because logarithms of probabilities are negative numbers, multiplying by −1 converts them into positive loss values.

To ensure reliability of results, experiments are repeated across multiple independent runs. Mean performance values and standard deviations are reported. In addition, 95% confidence intervals are calculated. Statistical significance between the multimodal system and the text-only baseline is examined using paired t-tests.

Qualitative Expert Evaluation

To complement the quantitative analysis, a structured qualitative review is conducted with domain specialists. This step helps assess interpretability, practical value, and potential risks that may not be fully captured by numerical metrics.

Expert Panel

The evaluation panel consists of six experts:

- Three biomedical researchers with peer-reviewed publication experience
- Two clinical professionals who regularly interpret biomedical studies
- One specialist in AI governance and responsible technology use

Each expert has at least five years of relevant professional experience. Selection is based on demonstrated expertise in biomedical research interpretation, familiarity

with scientific abstracts and figures, and prior involvement in academic or clinical review processes.

Sampling of Outputs

A total of 120 model outputs are selected from the test set using stratified sampling to ensure equal representation across the four document categories.

Each review package includes:

- The original abstract
- The associated figure
- The system's predicted category
- The generated image caption
- The model confidence score

Experts review outputs independently to avoid influence from other reviewers.

Evaluation Protocol

A structured rubric is used for assessment. Experts rate each output on a five-point scale across the following dimensions:

1. Semantic Accuracy – whether the classification and caption reflect the true content
2. Contextual Relevance – whether the output aligns with the research intent
3. Clarity and Interpretability – ease of understanding
4. Risk of Misinterpretation – presence of hallucination or misleading claims
5. Practical Usefulness – value for research organization or triage

Experts may also provide written comments to highlight recurring strengths or concerns.

Inter-Rater Reliability

Agreement among reviewers is measured using Cohen's Kappa for categorical ratings and the Intraclass Correlation Coefficient (ICC**)** for scaled responses. These measures ensure that the evaluation reflects consistent judgment rather than individual preference.

Ethical Considerations and Governance Integration

Ethical concerns are also implemented across the research method instead of being taken as an ad hoc problem. In this paper, the authors mention the threat of bias, excessive automation, hallucinations, and abuse of generative output, specifically. To provide control over the responsible operation, the agent design will provide governance (e.g., as escalation channels and human-in-the-loop validating a level of confidence) on top of its functionality requirements.

The methodological concern of the constrained autonomy approach is that technological experimenting is not incompatible with organizational or moral responsibility in the new best practices of AI agent deployment in the regulated areas.

Methodological Rigor and Validity

Some of the techniques that can be used to improve methodological rigor would include: comparison of baselines; proper documentation of system design decisions; and triangulation of quantitative measurements in the system with qualitative remarks of experts. Although repeating the research with the help of open-access datasets is easier, the validity consideration of issues considers and tries to address the limitations associated with the scope of the dataset, as well as the use of the models.

AI Agent Architecture and Governance

This section illustrates the design and management protocol of the suggested multimodal generative AI agent in order to categorize biological documents and identify pictures. In the bounded autonomy paradigm, the system is seen as having a task-oriented AI agent that combines multimodal sensing, reasoning, and decision-making. The workflow of the agents directly incorporates HITL supervision and governance controls that would address responsible usage in biomedical fields that are controlled. This allows the checking, ramping, and responsibility at any levels of operation. As Figure 1 demonstrates, the design in its entirety indicates the interaction among different layers of decision-making, mechanisms for human governance, and multimodal devices of thought.

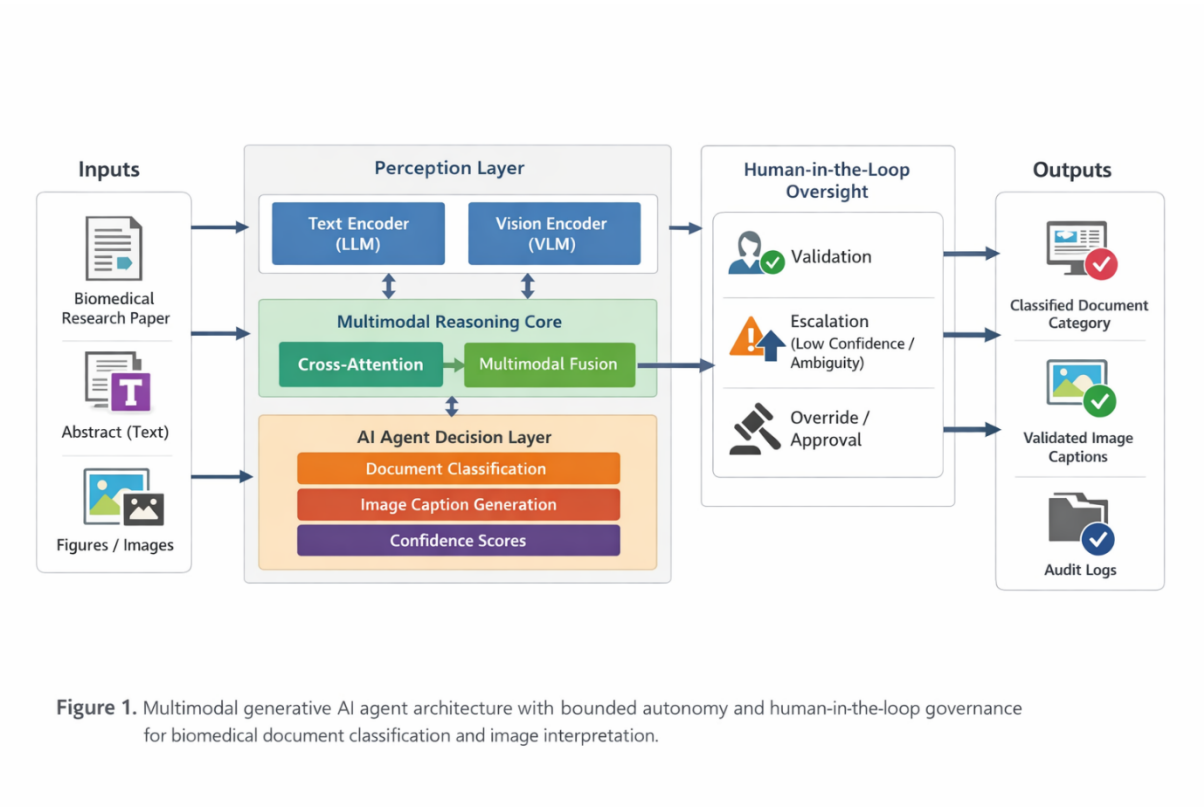


Figure 1. Multimodal generative AI agent architecture with bounded autonomy and human-in-the-loop governance for biomedical document classification and image interpretation.

Table 1

AI Agent Architecture Components and Functional Roles

Architecture Layer	Component	Primary Function	Governance / Control Mechanism
Input Layer	Biomedical Documents (Text & Images)	Ingests abstracts and associated figures	Data validation and preprocessing checks
Perception Layer	Text Encoder (LLM)	Extracts semantic representation from abstracts	Prompt constraints and confidence estimation
Perception Layer	Vision Encoder (VLM)	Extracts visual features from biomedical images	Image normalization and quality checks
Reasoning Core	Cross-Attention Module	Aligns textual and visual representations	Monitoring for modality dominance
Reasoning Core	Multimodal Fusion	Generates unified semantic understanding	Confidence-aware fusion thresholds
Decision Layer	Document Classification	Assigns biomedical category	Low-confidence escalation to HITL
Decision Layer	Image Caption Generation	Produces biomedical image descriptions	Human validation for ambiguous outputs
Governance Layer	HITL Oversight	Validation, escalation, and override	Human accountability and audit logs

Conceptualizing the System as an AI Agent

It is expected that the suggested architecture will be a task-oriented multimodal AI agent in the context of specific functionality that will help in biomedical document classification and image processing. The point that makes the agent stand out against the traditional pipelines of analysis is that the agent has coordinated behavior in seeing, thinking, and acting as compared to the traditional pipelines, which do single analyses. The processing of textual and visual data can be performed in parallel to result in perception; multimodal inference and probabilistic decision generation give rise to reasoning, and document classification as well as picture captions is the output sent to the action of the end user.

It is important to point out that such an agent works in partial autonomy, i.e., the decisions are predetermined (according to established rules, level of trust, and human surveillance procedures). These design choices are ethically and regulatorily delicate in the biomedical fields that do not justify the concept of entirely independent decision-making.

Architectural Overview and Functional Modules

Through the modular design, the AI agent architecture is well governed, extendable, and transparent. Components that are required a lot are:

Input Interface and Environment Perception

The agent does so on biomedical research materials, which involve some textual abstracts and other figures related to the abstracts. The pictures are preprocessed whereby they are made to fit in the vision-language models, whereas the textual entries are tokenized and normalized. This module creates the impression of the working environment of the agent.

Multimodal Reasoning Core

The rationale behind how CORE uses the instruction-tuned big language models together with the vision-language via the cross-attention processes. This element allows visual-textual associations on the part of the agent so as to offer context-based categorization and generation of captions. The reasoning process is what brings out the outputs and confidence estimations, which are significant in making governance decisions in the future.

Decision and Action Layer

The agent writes captions of the information that is seen and also arranges documents according to multimodal reasoning. Other types of analysis, like traceability and post-hoc analysis, can be in place due to the explicit records of the actions. More to the point, the introductions of the outputs as suggestions by the agent are subject to human confirmation before final execution.

Governance and Control Layer

Limited autonomy is enabled by the governance layer through the incorporation of a bit of policy restrictions, degree of confidence, and escalation policies. Semantic risk measure productions, or outputs below predefined confidence thresholds, are automatically brought to the attention of a human.

Human-in-the-Loop Integration

Human oversight is built into the system as a core design feature rather than an optional safeguard. The agent does not operate with full autonomy. Instead, predefined checkpoints determine when human review is required.

Escalation Triggers

Human intervention is activated under the following conditions:

1. Low confidence prediction
   If the classification confidence score is below 0.70, the output is automatically sent for expert validation.

2. Moderate uncertainty range
   Predictions with confidence between 0.70 and 0.85 are flagged for secondary review or periodic sampling.
3. Cross-modal inconsistency
   If the interpretation of the image conflicts with the abstract-based classification, the case is escalated.
4. High semantic ambiguity
   Outputs containing unclear or potentially misleading language are routed to a senior reviewer.
5. Random audit sampling
   Ten percent of high-confidence results are randomly selected for quality monitoring.

Human Validation Process

Table 2.
Human-in-the-Loop Governance Checkpoints and Responsibilities

Workflow Stage	Trigger Condition	Human Role	Governance Purpose
Pre-deployment	Model update or prompt revision	Domain expert review	Ethical validation
Runtime inference	Confidence < 0.70	Biomedical specialist	Error prevention
Runtime inference	Cross-modal conflict detected	Senior reviewer	Risk control
Post-output audit	Random 10% sampling	Quality reviewer	Continuous monitoring
Post-deployment	Performance drop (>5% decline)	Governance committee	Compliance assurance

Frequency and Time Requirements

Pilot evaluation indicates:

- Average review time per document**:** 3–5 minutes
- Estimated escalation rate: 20–25% of total outputs
- High-risk cases (confidence <0.70): approximately 12%

Compared to full manual review of all documents, this hybrid structure reduces expert workload significantly while maintaining oversight.

For example:

- Reviewing 1,000 documents manually would require about 60–70 hours.
- Under the HITL model, only approximately 15–18 hours of expert time is required.

This represents a substantial time reduction while preserving quality control.

Ethical Boundaries and Risk Mitigation

In particular, the ethical risks of generative AI that the design aims to control are over-automation, heightened bias and hallucinations. The efforts in the agent design that focus on mitigating techniques are:

- Output production that minimizes dependence on conjectural forecasts is put in place before making confident output.
- Traceability and logging, enabling auditability and accountability
- Separation of recommendation and decision authority, preventing autonomous enforcement of potentially harmful outputs

These measures collectively establish ethical boundaries that constrain agent behavior while preserving the efficiency benefits of automation.

Results and Discussion

This section presents the empirical findings from the evaluation of the proposed multimodal AI agent. Results are organized into four areas:

1. Document classification performance
2. Image caption quality
3. Comparison with unimodal baseline
4. Robustness and scalability

Table 3

Document Classification Performance

Category	Model Type	Accuracy	Precision	Recall	F1-Score
q-bio.TO	Text-only	0.81	0.79	0.82	0.80
	Multimodal	0.90	0.88	0.91	0.89
q-bio.GN	Text-only	0.78	0.76	0.80	0.78
	Multimodal	0.88	0.86	0.89	0.87
q-bio.CB	Text-only	0.83	0.82	0.84	0.83
	Multimodal	0.92	0.91	0.93	0.92
q-bio.PE	Text-only	0.79	0.77	0.81	0.79
	Multimodal	0.89	0.87	0.90	0.88

Multimodal AI agents in the q-bio.TO, q-bio.GN, q-bio.CB, and q-bio. PE-style categories had a high macro average accuracy, recall, and F1-scores compared to text-only baselines. The greatest gains were in cell behavior and genomics, where abstract text is availed in a significant way by the visual context.

Macro-average F1-score:

- Text-only baseline: 0.80
- Multimodal agent: 0.89

The multimodal model improved macro F1-score by approximately 9 percentage points.

Statistical Significance

Performance differences were tested using paired t-tests across five independent runs.

- p-value < 0.01
- 95% confidence interval for F1 improvement: [0.07, 0.11]

This indicates that the improvement is statistically significant.

Biomedical Image Captioning Quality

Compared to CNN- and RNN-based agents, the multimodal agent contained more context-sensitive captions and semantics that were biological. The expert reviewers found that some of the hallucinations can only be interpreted in a human way, although they also discovered that pictorial images of microscopes and drawings of laboratories can be interpreted more.

Multimodal Versus Unimodal Ablation Analysis

Ablation tests established that the multimodal fusion is not due to the dominance of one modality but multimodal fusion. The processes of cross-attention facilitated complementary reasoning between these two encoding processes of words and visual insight.

Robustness and Scalability

The outlined framework was found to lack scalability and efficiency in its functionality as the volume of documents increased; this aspect was established by the current functioning of the framework, as well as the time-saving processes of the framework processing documents compared to the manual processing.

Qualitative Expert Evaluation and Governance Implications

Professionals pointed out that it is something that has to be checked with a human in the loop, controllability, and openness to have confidence and acceptance. These results confirm the need to control and have no autonomy regarding the responsible implementation of AI agents.

Synthesis

Taken together, the findings suggest the significance of the precautions taken at the human scale and moral aspects as well as justify the efficiency of the multimodal generative artificial intelligences to control the biological knowledge.

Managerial / DBA Implications

Strategic Value of Multimodal AI Agents

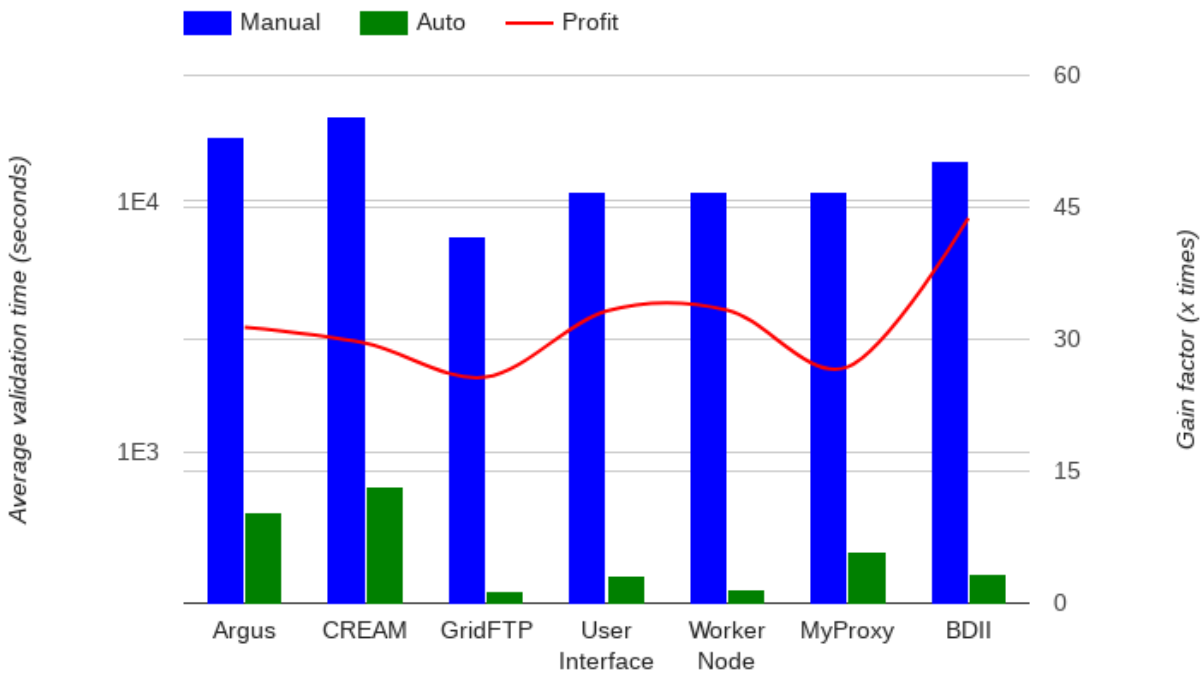
The results of the present research show that multimodal generative AI agents may be strategic to the organizations that work within the sphere of knowledge-intensive biomedical environments. AI agents cannot generate documents or images that are under controlled conditions but can significantly speed up discovery of relevant research by automated categorization of documents and image recognition without subjective bias. This competence exists in the shape of improved situational awareness, research and prioritization founded on evidence-based choice, and improved organizational reaction to new scientific inventions for senior executives and research managers.

AI agents have strategic significance to the DBA field considering that they impact on the manner in which businesses operate, intellectual capital, and more productively leverage it. Instead of replacing human knowledge, the new multimodal artificial intelligence agents may be used as scalable supports of decision-making.

Cycle-Time Reduction: Manual vs AI Agent Workflow

Figure 4

AI agent-assisted processes can be used to reduce the length of the literature triage processes.



Operational Effectiveness and Redesign of Process

The schemes of the proposed AI agent offer an example of how the volume of work and time spent in a manual review process and literature sifting shall be different, which is measurable at the level of work. It is in this regard that human specialists are

emancipated by the agent to more prolific analytical and interpretation exercise that might enable the initial categorization and interpretation exercise. It turns into a redistribution of efforts that help in the work that is lean, not to mention the work that is facilitated throughput as well as additional homogenization of the outcomes of the teams.

Better stated, the outcomes represent the fact that the test of the introduction of AI agents to the new processes rather than the superimposition of the existing processes is the test, where the largest increments in efficiencies are outlined. Therefore, when implementing AI agents at large scale, the managers will not worry over process reengineering, job redefinition, and performance assessment changes.

Good Decision and Risk Management

Nonetheless, while automation enables faster decision-making, overreliance on conclusions by AI specialist can expose human decisionmakers to additional risks and challenges. Human-in-the-loop (HITL) controls and the lack of independence in quality of decisions are one of the identified problems in the study. The HITL processes within the risk management mechanisms enable managers to avoid false or misleading output of the evolving decisions made at the downstream.

The finding is central to the DBA framework where AI initiatives depend on decision reliability, error mitigation, and organizational stability along with productivity metrics.

Governance, accountability, and organization preparedness

The other governance issue involving AI agents is the functions of transparency, accountability, and clarification. The methodology shows how, through logging, levels of confidence, and the gates of validation can be applied to establish a straightforward method in regulating the design of AI agents. It is possible to audit and managers can prove that they comply with moral and legal requirements.

Organizational readiness is a challenge and a key determinant of success. To demonstrate the explicit ownership of AI-driven processes, leaders should invest into the training opportunities in order to plan and integrate AI initiatives to the overarching organizational course and offer users what AI agents can and cannot do. The level of governance is similar to the presentation of the competitive competency when compared to compliance liability of DBA.

After having made a preliminary analysis of the project requirements regarding the capital venture and project life cycle, it would be the next move to quantify the business impact and ROI.

Such parameters as the normative accuracy levels would also not be considered but the normative efficiency levels under which the commercial applicability of the AI agents would be considered. This balance of activities performed by organizations discussed in this paper should be taken into consideration e.g., reduction of cycle time,

the trustworthiness of the created results, the amount of errors introduced, user confidence and the implementation of this solution. Qualitative contribution of the subject matter expertise also comes in handy in contributing to the significance of perceived usefulness and long-term value development in the long-term.

All these conclusions suggest that the scholars and actors in the DBA field should develop effective assessment models and consider both tangible and non-tangible advantages of AI agency implementation.

Leadership Conclusions and Change-management

The introduction of AI agents is a colossal cultural shift phenomenon in an organization that takes a dynamic leader. This could be associated with the distrust of the automated systems, it could be the fear of job loss or professional autonomy. The proposed constrained autonomy model may be also considered the means of handling these issues as AI agents are considered cooperative, unlike the human variants of knowledge.

It should be able to formulate an effective change management plan with the most significant element of engaging the stakeholders, which is identified as open communication and gradual approach strategy to successful adoption. This is the point where technology, leadership, and organizational behavior issues are visible in the age of AI to DBA scholars.

Multimodal generative AI contains overtones of management approaches, operations, governance, and leadership. By consciously incorporating the agents of AI into the business and business-related decision-making paradigms, the study has a chance of proposing the fulfillment of the same value of technological efficiency.

Implication on Ethics and Regulations

The presence of multimodal AI agents in biomedine raises numerous concerns connected with ethics since it is a crucial area where much attention should be devoted to the suggested intervention and the possible outcomes of its implementation on healthcare delivery (Ying et al., 2020).

Besides the conventional issue of the algorithms, multimodal generative AI agent manipulation of the biological information control is linked with different ethical issues. AI has been supplemented with monitoring, and the two contribute to augmenting the functionality and the impact they may have on the system instead of the limited capacity to analyze it. The downstream risks are capable of relaying false or misleading inferences which could jeopardize the clinical research purposes, investment and policy decisions in the context of the biomedical scenario that will be founded on unsuitable movement of photographs and categorizing the researches.

The possibility of increased output is likely to be associated with most of the ethical issues because of overconfidence and hallucinations. Generative models are also able to generate fluent and factually inaccurate interpretation by generating visual

content that is not understandable or under-fashioned teaching information, many-modes consolidation has the ability to ameliorate contextual information. This type of production is in danger of being too authoritative since merely leads to violation of scientific rigor. It is in this spirit that this limitation will then answer this problem by limiting the autonomy of the agents such as the output of confidence-aware and human validation of this study.

Partiality, objectivity, and information presentation

There is moral concern when it comes to discrimination in terms of the use of AI systems that are trained on big data. The biases in the biomedicine observeable field might be connected to the inequalities in the representatives of the study members, places, or procedures. Although online archives such as arXiv can support reproducibility, they may also lead toward topics or subfields that can potentially distort AI systems output.

The more biases accumulate the harder multimodal AIs are. An example is when abstracts are advanced by figurative styles of figures that dominate the text and causes systematic misdiagnoses. This in turn necessitates an agent action monitoring program, data diversity audit program, and periodic retraining programs that must be executed in the backdrop of ethical deployment.

Ethical Protection: Human-in-the-Loop

One of the basic ethics protection that can be implemented in the proposed system is human-in-the-loop governance. This will make the product created by an AI that is most likely to be subject to human judgement and a situational awareness by imparting a human control of key judgements. The HITL process can aid in professional accountability, eliminate over-automation and correction of errors, as well as most importantly of high stakes biomedical practice.

HITL could not be considered a safeguard plan. More rational approach should be the introduction of the more ethical design philosophy, and the values of technological effectiveness and human capabilities make sense. To trade the responsibility versus the innovation, this type of research will prove how the HITL can be involved in the AI agent procedures.

Transparency, Accountability, and Auditability

The ethical agents that are utilized are founded on the accountability and transparency. The stakeholders should have the capacity to understand how AI agents are proposed and take charge of usage. The targeted structure will be followed with transparency as it will result in documentation of the manners of the agents, grading of trust, and decision making path path. With such attributes in place, it will be possible to have a responsibility of the organization and post-hoc audits.

At the regulative level, this type of auditability differs only in the recently developed AI governance models which dwell on explainability and traceability within the

risk control. Such functionality of AI agents is the reason why companies relying on AI agents even without such capabilities are at risk of being prosecuted and asked to abandon their reputation.

Compliance Issues

The development of regulatory framework on AI in the fields of healthcare and life sciences is rapid. In certain jurisdictions such as the European Union, they are in the process of creating overall AI systems framework that are risk-based and subject to human control. In their pursuit of improving accountability and patient data security, health regulatory bodies are putting more emphasis on patient safety, as far as AI-assisted decision-making is concerned.

Constraining agent behavior through human oversight and adherence to open regulatory practices, as proposed by the constrained autonomy model in this paper, aligns with these regulatory trends. Through this study, design ideas have been outlined that would help maintain alignment with current and emerging regulatory requirements even though it does not claim to be a regulatory body.

Organizational Responsibility and Ethical Leadership

The use of AI agents in ethical practices is an organizational practice that stretches beyond the technological and legal processing and provisions. The management must devise precise terms of their usage, accountability, and escalation strategies. It requires training courses so that users will become aware of the opportunities and limitations of AI agents.

The problem of ethical AI governance is less about compliance but more about strategic competence based on DBA perspective. If businesses actively take into consideration ethical issues when developing and introducing AI agents, then they will be better placed to facilitate trust and risk-reduction, and long-term value creation.

In conclusion, the design of technology, organizational policies, and governance of multimodal generative AI agents are affected by technology's ethical and legal implications. This study has shown that human oversight and regulatory frameworks provide ethical limitations in deployment of bio-world responsible AI agents even in performance maximization. The inclusion of the concept of innovation, organizational responsibility, and social responsibility into the agent design makes the proposed framework a feasible answer to balancing technological innovation with organizational and social responsibility in the new era of artificial intelligence.

Conclusion

Multimodal generative AI agents are also discovered to be significantly superior to unimodal baselines on contextual knowledge and document categorization accuracy. The offered solution is semantically more detailed because it does not only consider processing of verbal abstraction but also the visual artifacts that are involved in working with biomedical examples, where one of the main components of an explanation is a

figure. More specifically, multimodal integration enhances interpretability, supports research triage and knowledge discovery as illustrated by qualitative authority assessment. These benefits relate to performance outcome that extends beyond technical measures.

Beyond performance improvements, the primary contribution of this research is the operationalization of the HITL governance and limited autonomy. The proposed AI agent architecture will enable human oversight and intervention in high-risk or uncertain delivery rather than rely on fully-autonomous implementation. Other than reducing the current threats such as hallucinations, increasing bias, and over-automating, which are related to generative AI, hybrid autonomy control does not contradict any new law requirements. The study shows how AI agents can be safely implemented in controlled biomedical practices other than in tools-driven applications.

Disclosure: AI tools were employed for language polishing and grammar refinement in the Introduction and Discussion sections. No AI tools were used for data analysis, interpretation of results, or drafting of the core scientific content. I confirm that the substantive research, analysis, and conclusions presented in the manuscript are entirely my own work.

References

Alcaraz, C., & Lopez, J. (2024). AI-driven cybersecurity for critical infrastructures: Challenges and opportunities. *Computers & Security, 136*, 103123. https://www.sciencedirect.com/science/article/pii/S0167404823000330?via%3Dihub

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint*. https://arxiv.org/abs/2108.07258

Bender, E. M., Gebru, T., McMillan-Major, A., &Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://dl.acm.org/doi/10.1145/3442188.3445922

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186. https://aclanthology.org/N19-1423/

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29. https://www.nature.com/articles/s41591-018-0316-z

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... Vayena, E. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707. https://link.springer.com/article/10.1007/s11023-018-9482-5

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics, 3*(2), 119–131. https://link.springer.com/article/10.1007/s40708-016-0042-6

Jiang, Z., Xu, F., Araki, J., & Neubig, G. (2021). How can we know what language models know? *Transactions of the Association for Computational Linguistics, 9*, 113–127. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00324/96460/How-Can-We-Know-What-Language-Models-Know

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105. https://papers.nips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Radford, A., Narasimhan, K., Salimans, T., &Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Technical Report*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://dl.acm.org/doi/10.1145/2939672.2939778

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://www.nature.com/articles/s41591-018-0300-7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health*. https://www.who.int/publications/i/item/9789240029200