scientific reports

OPEN

A scalable framework for evaluating multiple language models through cross-domain generation and hallucination detection

Sorup Chakraborty[1], Rajesh Chowdhury[1], Sourov Roy Shuvo[1], Rajdeep Chatterjee[1] & Satyabrata Roy[2]✉

Large language models (LLMs) have significantly advanced in recent years, greatly enhancing the capabilities of retrieval-augmented generation (RAG) systems. However, challenges such as semantic similarity, bias/sentiment, and hallucinations persist, especially in domain-specific applications. This paper introduces MultiLLM-Chatbot, a scalable RAG-based benchmarking framework designed to evaluate five popular LLMs GPT-4-Turbo, CLAUDE-3.7-Sonnet, LLAMA-3.3-70B, DeepSeek-R1-Zero, and Gemini-2.0-Flash across five domains: Agriculture, Biology, Economics, Internet of Things (IoT), and Medical. Fifty peer-reviewed research papers (10 per domain) were used to generate 250 standardized queries, resulting in 1,250 model responses. Texts from PDFs were extracted using PyPDF2, segmented to preserve factual coherence, embedded with sentence-transformer models, and indexed in Elasticsearch for efficient retrieval. Each response was analyzed across 4 dimensions: cosine similarity for semantic similarity, VADER sentiment analysis for sentiment detection, TF-IDF scoring, and named entity recognition (NER) for hallucination identification and factual verification. A composite scoring scheme aggregates these metrics to rank model performance. Experimental results show LLAMA-3.3-70B as the overall best-performing model, leading in all 5 domains. The proposed framework is implemented using Colab notebooks, which offer a reproducible, extensive pipeline for domain-specific LLM benchmarking. Through the combination of cross-domain analysis and multi-metric evaluation, this study fills in the gaps in current LLM benchmarking procedures and offers a modular architecture that can be adjusted to new domains and future LLM advancements. The findings inform model selection strategies for researchers and practitioners seeking trustworthy LLM deployment across diverse industrial and scientific sectors.

Keywords Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Semantic Similarity, Bias Detection, Hallucination Detection, Multi-Domain Benchmarking, Elasticsearch

The field of Natural Language Processing (NLP)[1] has gone through a dramatic transformation with the arrival of transformer-based architectures, beginning with the groundbreaking "Attention Is All You Need"[2] model. This innovation opened the way for the development of Large Language Models (LLMs)[3] such as GPT-4 Turbo, LLAMA-3.3-70B-Versatile, Claude-3.7 Sonnet, Gemini-2.0 Flash, and DeepSeek R1 Zero[4–8]. These models have demonstrated exceptional capabilities in diverse NLP tasks like text generation, summarization, and comprehension. However, despite their impressive performance, LLMs continue to face significant challenges, particularly in areas such as semantic similarity, bias, and hallucination[9–11].. These issues reduce the reliability, factual accuracy, and ethical soundness of model outputs, highlighting the need for comprehensive evaluation frameworks. In response to these challenges, Retrieval-Augmented Generation (RAG)[12] has emerged as a promising solution to enhance model grounding by conditioning outputs on retrieved external documents. Benchmarks like HELM[13] and TruthfulQA[14] have made advances in assessing general factuality and performance

[1]School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar 751024, Odisha, India. [2]Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur 303007, Rajasthan, India. ✉email: satyabrata.roy@jaipur.manipal.edu

but fail to address domain-specific challenges and offer multidimensional evaluation frameworks that assess semantic similarity, sentiment bias, hallucination detection, and factual verification. To make up these gaps, we introduce MultiLLM-Chatbot, a scalable, domain-diverse framework designed for the comprehensive evaluation of RAG systems across multiple LLMs. This framework addresses key evaluation challenges by covering five distinct domains: Medicine, Internet of Things (IoT), Agriculture, Economics, and Biology. We developed a dataset of 50 peer-reviewed research papers and designed 250 standardized queries across 5 different LLMs resulting in 1250 responses targeting a range of cognitive tasks, such as comprehension, comparison, summarization, and inferential reasoning. This structured approach enables a detailed comparison of model performance across varied domains and tasks.

Despite the impressive progress in the development of Large Language Models (LLMs), their deployment in real-world, domain-specific settings continues to be hampered by several persistent challenges most notably, hallucinated outputs, semantic drift, and underlying biases. These shortcomings become particularly problematic in high-stakes domains such as healthcare, finance, and scientific research, where accuracy, reliability, and contextual relevance are critical. Although current benchmarking tools provide general evaluations of language model performance, they often fail to capture the intricacies of domain-specific reasoning or support multidimensional assessment. While many evaluations focus on fluency and superficial accuracy, what remains lacking is a comprehensive framework that rigorously examines deeper dimensions such as semantic coherence, factual accuracy, and logic grounded in domain expertise. Our research addresses this gap by proposing a scalable, robust evaluation methodology grounded in Retrieval-Augmented Generation (RAG). By anchoring model outputs in authoritative sources, this approach enables more meaningful and context-aware comparisons across LLMs, especially within specialized fields.

Our methodology integrates a multi-model benchmarking approach, generating 1,250 responses by querying five leading LLMs under controlled conditions. To ensure thorough evaluation, we combine semantic similarity scoring, sentiment bias analysis, hallucination detection through TF-IDF, and Named Entity Recognition (NER)-driven factual verification. The results demonstrate that LLAMA-3.3-70B-Versatile outperforms other models in all domains.

To address existing gaps in LLM evaluation, this work introduces MultiLLM-Chatbot a comprehensive, domain-specific benchmarking framework that evaluates LLMs across multiple dimensions, including semantic similarity, sentiment bias, factual consistency, and hallucination detection within a retrieval-augmented generation (RAG) setup. Unlike prior benchmarks limited to single domains or isolated metrics, our framework integrates dense retrieval, multi-model comparison, and a composite evaluation scheme to offer a holistic assessment. We have conducted extensive experiments using a large dataset comprising 50 research articles, 250 queries, and 1250 generated responses across five diverse fields such as Agriculture, Biology, Economics, IoT, and Medicine. This framework provides scalable, reproducible insights for reliable real-world deployment.

Related work

Recent progress in large language models (LLMs) and retrieval-augmented generation (RAG) systems has significantly advanced information retrieval and text generation research. Early RAG frameworks like REALM[15] and RAG[16] pioneered the integration of dense retrieval with generative models to improve factual grounding. Subsequent enhancements in retrieval methods, such as ColBERT[17] and hybrid approaches combining BM25 with dense retrieval[18], further improved retrieval quality. However, evaluations have largely been confined to single-domain settings, limiting insights into cross-domain performance. Semantic retrieval has increasingly relied on sentence-level embeddings, with models like Sentence-BERT[19] becoming standard for dense retrieval tasks. Benchmark datasets such as MS MARCO[20] and BEIR[21] emphasize evaluating retrieval systems across diverse tasks but often neglect generation quality assessment.

The issues of bias and sentiment in LLM-generated outputs have drawn considerable attention, with studies highlighting how LLMs can propagate biases related to gender, politics, and social contexts[22,23]. Lightweight tools like VADER[24] enable large-scale sentiment analysis, while benchmarks such as HELM[25] incorporate bias and fairness assessments. However, domain-specific evaluations of bias in RAG workflows remain underexplored. Recent work by Awlla and Kozhin[26] on sentiment analysis in low-resource Central Kurdish demonstrates the importance of adapting LLM-based frameworks to linguistic and cultural diversity. Similarly, advancements in Named Entity Recognition (NER) for low-resource languages, such as Politov et al.'s[27] transfer learning approach for Slavic languages, improve factuality and hallucination detection in information-scarce environments.

Detecting hallucinations in model outputs remains a significant challenge, with efforts like FRANK[28] and FactCC[29] focusing primarily on summarization tasks and lacking multi-domain RAG applications. Techniques such as TF-IDF-based divergence measures and entity-level fact-checking using NER[30] show promise but require broader cross-domain validation. Multi-model evaluation frameworks like MMLU[31] and BIG-bench[32] assess reasoning capabilities but do not specifically address retrieval-grounded generation. While BEIR benchmarks cross-domain retrieval, it omits generation quality and consistency checks. To address these gaps, our work introduces MultiLLM-Chatbot, a unified framework combining dense retrieval, multi-model generation, sentiment and hallucination analysis, and comprehensive multi-domain benchmarking, setting a new standard for LLM evaluation.

Methodology

This section shows the complete workflow of the MultiLLM-Chatbot framework, encompassing data preparation, model integration, retrieval infrastructure, response generation, and multi-dimensional evaluation. We designed each phase to prioritize fairness, reproducibility, and coverage across diverse domains. Additionally, our project follows the step-by-step process outlined in Algorithm 1.

Data collection

The evaluation framework was built upon a carefully curated dataset spanning five distinct fields: Medical, Internet of Things (IoT), Agriculture, Economics, and Biology. Ten peer-reviewed research articles were selected per domain, resulting in a collection of 50 papers. Selection criteria included publication quality, topical relevance, technical depth, and the availability of clean, machine-readable PDFs. From each paper, five standardized queries were systematically formulated, resulting in a total of 250 queries. These questions were designed to assess a wide spectrum of cognitive skills, including factual recall, inference, summarization, and comparative reasoning. Corresponding ground-truth answers were extracted directly from the source papers, forming a reliable benchmark for subsequent evaluations. The structured and domain-diverse nature of this dataset ensures equitable and meaningful model comparisons across specialized areas of knowledge. Our observations indicate that the evaluated models are primarily trained in English[33] and show reduced performance on less-resourced or non-English languages. Therefore, for this study, we chose to conduct all evaluations in English, with future work aimed at expanding coverage to a broader range of languages for more inclusive assessments.

LLM integration

Our evaluation framework integrates five cutting-edge large language models: GPT-4 Turbo, Claude 3.7 Sonnet, Gemini 2.0 Flash, LLAMA-3.3-70B Versatile, and DeepSeek R1 Zero each chosen for their unique strengths to address the diverse demands of knowledge-intensive tasks. GPT-4 Turbo contributes advanced reasoning and summarization, enabling context-rich, coherent outputs. Claude 3.7 Sonnet enhances factual accuracy and safety, reducing hallucinations in critical applications. Gemini 2.0 Flash offers fast inference for real-time and latency-sensitive scenarios. LLAMA-3.3-70B Versatile brings open-weight adaptability and strong generalization, supporting flexible fine-tuning across varied domains. DeepSeek R1 Zero, optimized for retrieval-augmented generation, ensures precise, fact-grounded responses. Collectively, these models form a robust foundation for comprehensive evaluation across reasoning, speed, factual reliability, and domain versatility.

Vector search and retrieval

The RAG pipeline commenced with text extraction from PDFs using PyPDF2. Cleaning processes removed irrelevant elements such as headers, footers, page numbers, and extraneous metadata, ensuring a high-quality retrieval corpus.

The MultiLLM-Chatbot QA Pipeline begins with the ingestion of PDF documents and user queries. Initially, the system uploads and parses each page of the input PDFs to extract text, which is stored along with associated metadata. These documents are then segmented into overlapping text chunks to preserve contextual continuity. Each chunk is embedded using SentenceTransformer models, transforming the text into dense vector representations. The resulting embeddings are indexed and stored in Elasticsearch for efficient retrieval. During runtime, the system enters an interactive loop, awaiting user input in the form of questions. Upon receiving a query, the system retrieves the top-k most semantically similar chunks from the Elasticsearch index using vector search techniques. These retrieved chunks form the context used to build a coherent response prompt. The constructed prompt, consisting of the user query and relevant context, is then sequentially passed to each integrated large language model (LLM), specifically Llama, Gemini, GPT-4, Claude, and DeepSeek. Each LLM processes the prompt and generates its own response, which is subsequently stored. After all models have responded, their outputs are displayed to the user. This process repeats iteratively, allowing for multi-model comparison and evaluation on a per-query basis.

Extracted texts were segmented into coherent chunks of 200 to 300 words, maintaining sentence boundaries through spaCy's dependency parsing[34] to preserve logical flow. Additionally, a 20% overlap was introduced between consecutive chunks to enhance contextual continuity. Semantic embeddings were generated using the "all-MiniLM-L6-v2" model from Sentence Transformers[35], producing 384-dimensional normalized vectors for each chunk. These embeddings were indexed within an Elasticsearch[36] cluster configured for Approximate Nearest Neighbor (ANN) search via Hierarchical Navigable Small World (HNSW)[37] graphs, employing cosine similarity as the distance metric. This retrieval architecture ensured that LLMs received highly relevant, document-grounded contexts, reducing reliance on model memorization. The languages supported by the evaluated LLMs are summarized.

Response generation

Once the retrieval system and LLM access were set up, we generated responses following a consistent and standardized approach. For each queries:

- The top semantically relevant text chunks were retrieved from Elasticsearch.
- Retrieved contexts were concatenated and embedded into a standardized prompt template alongside the user query.
- The complete prompt was submitted to each of the five LLMs.

This system produced five distinct responses per query, resulting in a total of 1,250 responses. All outputs were stored in structured JSON files and aligned with the corresponding ground-truth references, providing a consistent and scalable basis for downstream evaluations.

Evaluation metrics

We used a multi-faceted evaluation framework to thoroughly assess how each model performed across four key areas. Instead of assigning weighted scores, we applied normalization techniques to make the comparisons fair and transparent.

Semantic similarity
Semantic similarity was measured using cosine similarity (Equation 1) between sentence embeddings of the model-generated response and the ground-truth answer, both computed via the *all-MiniLM-L6-v2* model. Higher scores indicated greater semantic alignment with the source material.

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{1}$$

Where:

- $A \cdot B$ = Dot product of vectors A (model response) and B (ground-truth text).
- $\|A\|$ = Magnitude (length) of vector A.
- $\|B\|$ = Magnitude (length) of vector B.

Bias detection (sentiment analysis)
Sentiment bias was assessed using the VADER (Equation 2) sentiment analysis tool. Since the source documents are academic and intended to be neutral, responses deviating toward strong positive or negative sentiments were considered biased.

$$\text{VADER's Compound Score} = \frac{\text{Sum of Sentiment Scores}}{\sqrt{(\text{Sum of Sentiment Scores})^2 + \alpha}} \tag{2}$$

Where:

- α is a small constant to smooth the score.
- "Sum of Sentiment Scores" = Total sentiment intensity after applying VADER's heuristics (e.g., intensifiers like "very", negations like "not good", punctuation emphasis like "!!!").

Hallucination detection
Two independent methods were employed to detect hallucinations:

- **NER-Based Fact Checking:** Named entities (persons, locations, dates, quantities) were extracted and compared against those in the ground-truth references, computing precision, recall, and F1 scores (Equation 3, 4 & 5). **Precision (P):**

$$P = \frac{\text{Correctly identified entities}}{\text{Total entities in LLM response}} \tag{3}$$

- **Recall (R):**

$$R = \frac{\text{Correctly identified entities}}{\text{Total entities in reference document}} \tag{4}$$

- **F1-Score:**

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{5}$$

- **TF-IDF Similarity:** Lexical overlap between the model response and retrieved context was evaluated using TF-IDF (Equation 6, 7 & 8) vectorization. Lower overlap scores were indicative of higher hallucination risks.

Term Frequency (TF)

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{w \in d} f_{w,d}} \tag{6}$$

Where:

- $f_{t,d}$ = Number of times term t appears in document d.
- $\sum_{w \in d} f_{w,d}$ = Total number of terms in document d.

Inverse Document Frequency (IDF)

$$\text{IDF}(t) = \log\left(\frac{N}{df_t}\right) \tag{7}$$

Where:

- N = Total number of documents (chunks).
- df_t = Number of documents where the term t appears.

TF-IDF

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t) \tag{8}$$

Overall scoring system
Instead of applying fixed weightings to each metric, the following normalization strategies were used:

- **Min-Max Normalization:** Scales scores within the [0, 1] range using Equation 9, preserving relative differences.

$$x_{\text{normalized}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{9}$$

Where:
- x = Original score.
- $x_{\min}$ and $x_{\max}$ = Minimum and maximum values in the column.

- **Z-Score Normalization:** Centers scores around the mean and scales them according to standard deviation, highlighting relative performance deviations using Equations 10 & 11.

$$z = \frac{X - \mu}{\sigma} \tag{10}$$

Where:
- X = Original score.
- μ = Mean of the metric.
- σ = Standard deviation of the metric.

Standard deviation (σ) is calculated as:

$$\sigma = \sqrt{\frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_n - \mu)^2}{n}} \tag{11}$$

Final model rankings were derived directly from normalized scores across all metrics, maintaining metric neutrality and ensuring statistically robust comparisons across diverse domains.

Novelty and research contributions

The novelty of this work lies in its comprehensive, domain-specific benchmarking framework that systematically evaluates LLMs across multiple dimensions semantic similarity, sentiment bias, factual consistency, and hallucination detection within retrieval-augmented generation (RAG) systems. Unlike prior benchmarks that focus on single-domain evaluations or isolated metrics, MultiLLM-Chatbot integrates dense retrieval (via Elasticsearch and sentence-transformers), multi-model comparison (GPT-4-Turbo, CLAUDE-3.7-Sonnet, LLAMA-3.3-70B, DeepSeek-R1-Zero, and Gemini-2.0-Flash), and a composite scoring scheme (combining cosine similarity, VADER sentiment analysis, TF-IDF divergence, and NER-based fact-checking) to provide a holistic assessment of LLM performance. The framework's modular design ensures reproducibility and scalability, enabling seamless adaptation to new domains and future LLM advancements. By leveraging peer-reviewed research papers across five diverse fields (Agriculture, Biology, Economics, IoT, and Medical), this study addresses critical gaps in cross-domain evaluation while offering actionable insights for reliable LLM deployment in real-world applications.

Proposed framework

The architecture of MultiLLM-Chatbot, as described in Figure 1 and in Algorithm 1, is designed as a modular, scalable pipeline to support comprehensive retrieval-augmented evaluations across multiple large language models (LLMs). The process starts with extracting clean, structured text from research PDFs. This text is divided into semantically coherent chunks and converted into high-dimensional embeddings using SentenceTransformer models. The resulting vectors are indexed in an Elasticsearch cluster for fast and accurate semantic retrieval during inference. During query execution, the user's question is embedded using the same model and used to retrieve the most relevant chunks from the document corpus. These retrieved chunks are combined with the query to construct a contextually grounded prompt, which is then forwarded to five integrated LLMs GPT-4 Turbo, Claude 3 Sonnet, Gemini Flash, LLaMA 3-70B Versatile, and DeepSeek R1 Zero via their respective APIs. The responses from each model are then collected and subjected to downstream evaluation tasks, including semantic similarity measurement, sentiment bias analysis, and hallucination detection using Named Entity Recognition (NER) and TF-IDF cross-checking.

To ensure consistency and comparability across models, outputs are normalized using both Min-Max and Z-score techniques. This pipeline architecture was chosen after reviewing multiple state-of-the-art approaches in LLM evaluation and RAG-based reasoning systems. Prior studies have demonstrated the effectiveness of

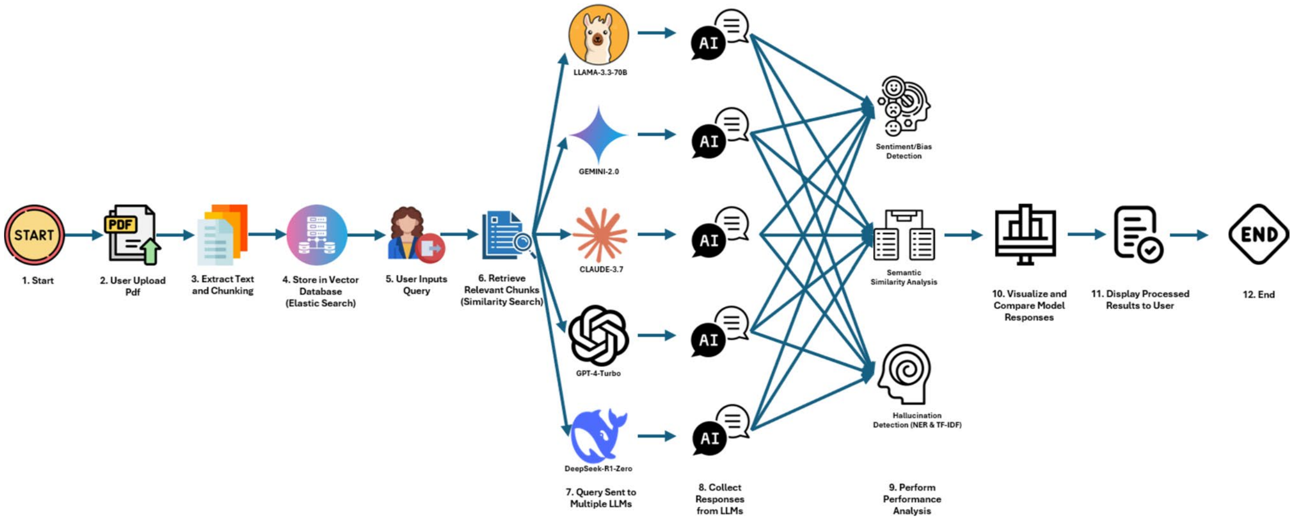
START
PDF
1. Start
2. User Upload Pdf
3. Extract Text and Chunking
4. Store in Vector Database (Elastic Search)
5. User Inputs Query
6. Retrieve Relevant Chunks (Similarity Search)
LLAMA-3.3-70B
GEMINI-2.0
CLAUDE-3.7
GPT-4-Turbo
DeepSeek-R1-Zero
7. Query Sent to Multiple LLMs
AI
8. Collect Responses from LLMs
Sentiment/Bias Detection
Semantic Similarity Analysis
Hallucination Detection (NER & TF-IDF)
9. Perform Performance Analysis
10. Visualize and Compare Model Responses
11. Display Processed Results to User
END
12. End

Fig. 1. Workflow architecture of the MultiLLM-Chatbot highlighting document processing, LLM-based response generation, and evaluation through semantic, sentiment, and hallucination metrics.

retrieval-augmented generation in reducing hallucinations and improving factual grounding in both open-domain and domain-specific applications (Lewis et al., 2020;). Moreover, recent works have highlighted the limitations of relying solely on end-to-end language model outputs without integrated retrieval or grounding layers particularly in scientific, biomedical, and legal domains where factual precision is non-negotiable (Khattab et al., 2022;)[38]. By structuring the system as a modular, multi-LLM comparison pipeline, we also address key evaluation challenges noted in prior research, such as semantic drift, hallucinated evidence, and domain-insensitive fluency scoring. The modularity allows rapid adaptation to new LLMs or domain-specific corpora, ensuring future extensibility. Furthermore, integrating

```
**Input:** PDF files $\mathcal{F}$, User queries $\mathcal{Q}$
**Output:** Evaluated responses $\mathcal{R}$ with metrics
**Phase 1: Document Processing**
$\mathcal{D} \leftarrow \emptyset$                                          // Initialize document collection
**foreach** $(filename, bytes) \in \mathcal{F}$ **do**
    Extract text $T$ from PDF bytes using PyPDF2
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(T, metadata)\}$                   // Store with source metadata
**end**
$\mathcal{C} \leftarrow$ SplitDocuments($\mathcal{D}$)                    // Chunk texts (1000 chars, 200 overlap)
**Phase 2: Vector Database Setup**
$\mathcal{E} \leftarrow$ SentenceTransformer("all-MiniLM-L6-v2")              // Initialize embeddings
$\mathcal{V} \leftarrow$ ElasticsearchStore($\mathcal{C}, \mathcal{E}$)                  // Index chunks in Elasticsearch
**Phase 3: Multi-LLM Query Processing**
$\mathcal{M} \leftarrow \{Llama, Gemini, OpenAI, Claude, DeepSeek\}$                  // Model endpoints
**foreach** $q \in \mathcal{Q}$ **do**
    $\mathcal{C}_q \leftarrow \mathcal{V}$.similarity_search($q, k = 5$)                  // Retrieve top 5 chunks
    $context \leftarrow$ concat($\mathcal{C}_q$)
    **foreach** $m \in \mathcal{M}$ **do**
        $r_m \leftarrow$ APIRequest($m$, $q$, $context$)                        // Parallel LLM queries
        $\mathcal{R}[q][m] \leftarrow r_m$
    **end**
**end**
**Phase 4: Response Evaluation**
Initialize metrics: $\mathcal{S}_{cos}, \mathcal{S}_{sent}, \mathcal{S}_{tfidf}, \mathcal{S}_{ner}$
**foreach** $q \in \mathcal{Q}$ **do**
    **foreach** $m \in \mathcal{M}$ **do**
        $\mathcal{S}_{cos}[q][m] \leftarrow \cos(\text{embed}(r_m), \text{embed}(\mathcal{C}_q))$
        $\mathcal{S}_{sent}[q][m] \leftarrow$ VADER($r_m$)                             // Sentiment polarity
        $\mathcal{S}_{tfidf}[q][m] \leftarrow$ TF-IDF($r_m, \mathcal{D}$)
        $\mathcal{S}_{ner}[q][m] \leftarrow \frac{|E(r_m) \cap E(\mathcal{D})|}{|E(r_m)|}$                    // NER entity match ratio
    **end**
**end**
**Phase 5: Visualization**
Generate comparative plots for $\{\mathcal{S}_{cos}, \mathcal{S}_{sent}, \mathcal{S}_{tfidf}, \mathcal{S}_{ner}\}$
**return** $\mathcal{R}, \{\mathcal{S}_{\ldots}\}$
```

Algorithm 1. MultiLLM Benchmarking Pipeline

multiple LLMs in a parallelized architecture facilitates comparative model auditing, consistent with best practices in robust AI evaluation.

The architecture of the MultiLLM-Chatbot, as depicted in Figure 1, is built as a modular and scalable framework to support multi-dimensional evaluation of large language models (LLMs) in a retrieval-augmented generation (RAG) setting. The pipeline begins with document ingestion, where research PDFs are parsed and converted into clean, structured text using PyPDF2. These documents are segmented into semantically coherent chunks of 1000 characters with a 200-character overlap, enabling better context handling during retrieval. SentenceTransformer models (specifically, `all-MiniLM-L6-v2`) are then used to generate embeddings for each chunk, which are stored in an Elasticsearch index for fast semantic search. This embedding-based retrieval enables accurate chunk selection based on user queries and grounds the LLM responses in domain-specific knowledge.

The query processing module initiates when a user submits a question. The query is embedded using the same SentenceTransformer model and compared with the indexed document chunks to retrieve the top-5 most relevant segments. These chunks are concatenated with the user query to form a grounded prompt, which is forwarded to five state-of-the-art LLMs: GPT-4 Turbo, Claude 3 Sonnet, Gemini 2.0 Flash, LLaMA 3-70B Versatile, and DeepSeek R1 Zero. These models are queried in parallel via their respective APIs, and their generated responses are collected for downstream evaluation. This multi-model interaction design enables robust comparison and auditing of model behavior under identical input contexts a key strength of the proposed system.

The evaluation phase of the pipeline applies a set of objective metrics to assess each model's output. Semantic similarity is measured using cosine similarity between the response and retrieved context embeddings. Sentiment polarity is analyzed using the VADER tool[24], which detects emotional bias. To evaluate factual consistency and hallucinations, we use a combination of TF-IDF divergence and Named Entity Recognition (NER)-based entity matching, comparing named entities in the response with those in the original source documents[30]. Finally, all evaluation scores are normalized using Min-Max and Z-score techniques to ensure consistency across models. This integrated scoring system is designed to address known limitations in LLM evaluation, including semantic drift, hallucinated content, and domain-insensitive fluency metrics[38].

Our modular architecture is informed by prior works in RAG systems[16,38] and LLM benchmarking[21], and is specifically tailored for real-world, cross-domain applications in fields such as Agriculture, Biology, Economics, IoT, and Medicine. By enabling transparent, side-by-side model auditing and flexible integration of new models or metrics, the MultiLLM-Chatbot framework aims to offer a more reliable and extensible foundation for future LLM research and deployment.

Results

Domain-specific analysis

The analysis of a sample query according to Figure 2, "What are the three main strategies incorporated into the Energy Management Scheme (EMS) proposed in EMS: An Energy Management Scheme for Green IoT Environments, and how does each address energy challenges in heterogeneous IoT nodes?"[39] reveals that Llama, Gemini, and Claude achieve high semantic similarity, with Llama and Gemini closely leading at a score of 0.92. Sentiment analysis across all models shows predominantly neutral outputs, with minimal emotional bias. In terms of factual consistency, Claude and Llama achieve the highest TF-IDF similarity scores, indicating strong alignment with the source material, whereas DeepSeek records the lowest, suggesting a higher rate of hallucination. Interestingly, DeepSeek performs best in NER-based factual accuracy, though Llama, Claude, and OpenAI also show strong results. Overall, Llama and Claude exhibit the best combined performance in terms of both semantic relevance and factual grounding. We compare the performance of each LLM within each of the five domains. Performance varies significantly between models, highlighting the importance of domain-specific LLM deployment strategies.

Agriculture

When evaluating all queries within the agriculture domain, aggregated results confirm that Llama and Claude lead in semantic similarity (0.857), with OpenAI following closely at 0.853, reflecting strong alignment with the reference answers as shown in Figure 3 and Table 1. Sentiment scores remain mostly neutral between models, and Gemini displays the highest neutral sentiment (0.910). Regarding the factual accuracy, Llama outperforms other models, achieving the highest TF-IDF similarity (0.453) and the NER-based entity recognition score

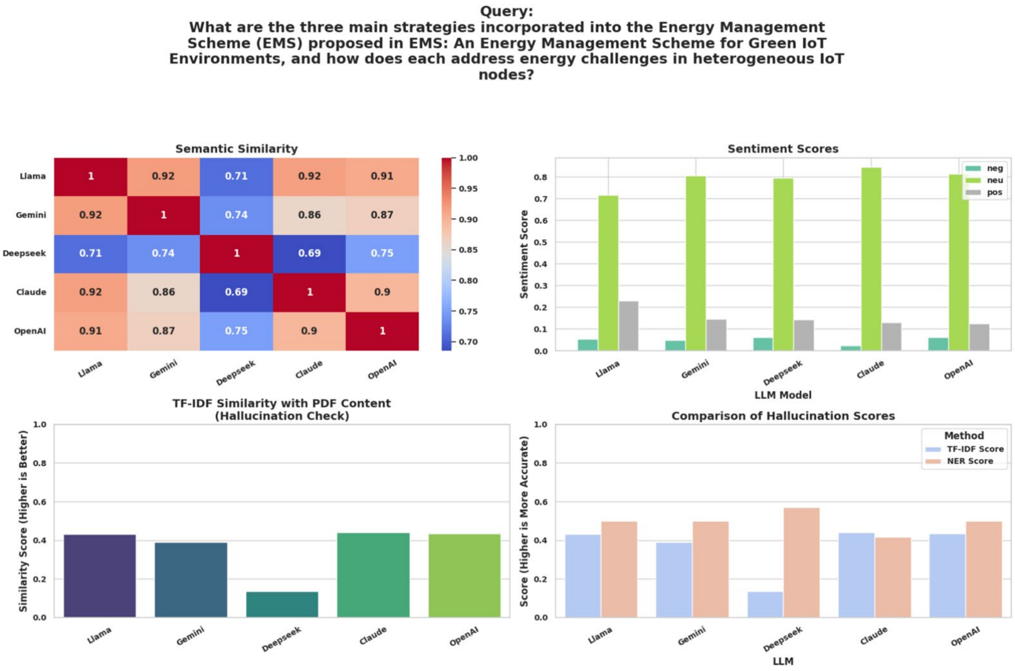

Fig. 2. This figure presents a multi-metric comparison of LLM performance for a IOT domain query. The top-left heatmap shows semantic similarity between model outputs, while the top-right bar chart illustrates sentiment distribution across responses. The bottom-left graph displays TF-IDF similarity with source content (for hallucination detection), and the bottom-right compares hallucination scores using both TF-IDF and NER methods.

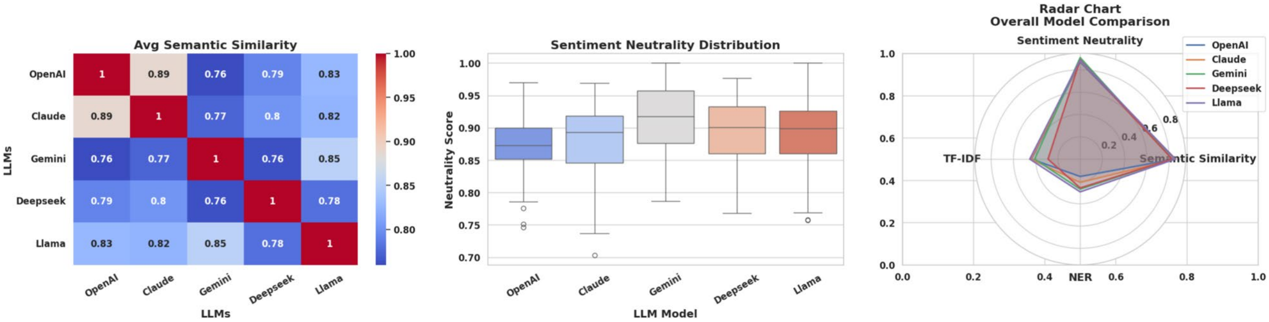

Fig. 3. This figure provides an aggregated overview of model-level performance. The left heatmap shows average semantic similarity between outputs of different LLMs, indicating alignment in understanding. The center box plot illustrates the distribution of sentiment neutrality scores, highlighting how balanced or biased the models' responses are. The right radar chart summarizes overall performance across key metrics semantic similarity, sentiment neutrality, TF-IDF, and NER accuracy enabling quick visual comparison of the models' strengths and weaknesses.

Domain	LLM Model	Semantic Similarity (Avg.)	Sentiment Neutrality (Avg.)	TF-IDF Similarity (Avg.)	NER-based Accuracy (Avg.)
Agriculture	Llama	0.857	0.888	0.453	0.294
	Gemini	0.830	0.910	0.409	0.270
	Deepseek	0.827	0.891	0.289	0.259
	Claude	0.857	0.882	0.442	0.209
	OpenAI	0.853	0.871	0.438	0.156
Biology	Llama	0.822	0.908	0.358	0.198
	Gemini	0.791	0.919	0.304	0.131
	Deepseek	0.803	0.897	0.278	0.163
	Claude	0.792	0.890	0.361	0.245
	OpenAI	0.814	0.897	0.300	0.143
Economics	Llama	0.761	0.861	0.426	0.205
	Gemini	0.728	0.861	0.413	0.189
	Deepseek	0.705	0.839	0.229	0.153
	Claude	0.701	0.855	0.371	0.166
	OpenAI	0.714	0.847	0.321	0.169
IOT	Llama	0.837	0.825	0.444	0.501
	Gemini	0.822	0.854	0.432	0.368
	Deepseek	0.796	0.829	0.210	0.375
	Claude	0.824	0.849	0.407	0.395
	OpenAI	0.832	0.829	0.338	0.416
Medical	Llama	0.841	0.855	0.411	0.209
	Gemini	0.831	0.871	0.394	0.197
	Deepseek	0.817	0.877	0.340	0.218
	Claude	0.775	0.865	0.345	0.233
	OpenAI	0.801	0.836	0.256	0.145

Table 1. This table presents a comparative evaluation of five large language models (Llama, Gemini, Deepseek, Claude, and OpenAI) across five domains Agriculture, Biology, Economics, IoT, and Medical using four key metrics: semantic similarity, sentiment neutrality, TF-IDF similarity, and NER-based accuracy. The results highlight model performance variations based on domain and metric, providing insights into their contextual strengths.

(0.294), suggesting excellent factual grounding. In contrast, DeepSeek records the lowest TF-IDF (0.289), and OpenAI records the lowest NER scores (0.156), indicating a higher tendency to hallucination. Gemini maintains steady performance across all metrics, balancing semantic understanding with factual reliability. Overall, Llama consistently outperforms others, with OpenAI showing strong semantic similarity but only moderate factual accuracy, while DeepSeek lags on most evaluation criteria. The final rankings within the agriculture domain place Llama firmly in the top position, ranking first in both min-max and z-score normalization methods as presented in Table 2. Gemini secures the second position with a balanced and strong performance in all metrics.

Domain	Model	MinMax Sem	MinMax Sent	MinMax TF-IDF	MinMax NER	MinMax Score	Z Sem	Z Sent	Z TF-IDF	Z NER	Z Score
Agriculture	Llama	1.000	0.436	1.000	1.000	3.436	0.909	−0.031	0.775	1.143	2.796
	Gemini	0.100	1.000	0.732	0.826	2.658	−1.103	1.690	0.046	0.657	1.290
	Claude	1.000	0.282	0.933	0.384	2.599	0.909	−0.501	0.593	−0.580	0.421
	OpenAI	0.867	0.000	0.909	0.000	1.775	0.611	−1.361	0.527	−1.654	−1.878
	Deepseek	0.000	0.513	0.000	0.746	1.259	−1.326	0.203	−1.941	0.434	−2.630
Biology	Llama	1.000	0.621	0.964	0.588	3.172	1.449	0.569	1.135	0.533	3.687
	Claude	0.032	0.000	1.000	1.000	2.032	−1.021	−1.198	1.225	1.671	0.677
	OpenAI	0.742	0.241	0.265	0.105	1.354	0.791	−0.510	−0.607	−0.799	−1.126
	Gemini	0.000	1.000	0.313	0.000	1.313	−1.104	1.649	−0.486	−1.090	−1.030
	Deepseek	0.387	0.241	0.000	0.281	0.909	−0.115	−0.510	−1.267	−0.315	−2.208
Economics	Llama	1.000	1.000	1.000	1.000	4.000	1.808	0.986	1.033	1.557	5.384
	Gemini	0.450	1.000	0.934	0.692	3.076	0.286	0.986	0.852	0.686	2.809
	Claude	0.000	0.727	0.721	0.250	1.698	−0.959	0.282	0.265	−0.566	−0.979
	OpenAI	0.217	0.364	0.467	0.308	1.355	−0.360	−0.657	−0.433	−0.403	−1.852
	Deepseek	0.067	0.000	0.000	0.000	0.067	−0.775	−1.596	−1.717	−1.274	−5.362
IoT	Llama	1.000	0.000	1.000	1.000	3.000	1.044	−1.028	0.901	1.875	2.792
	Gemini	0.634	1.000	0.949	0.000	2.583	−0.014	1.415	0.762	−0.896	1.268
	Claude	0.683	0.828	0.842	0.203	2.555	0.127	0.994	0.473	−0.333	1.260
	OpenAI	0.878	0.138	0.547	0.361	1.924	0.691	−0.691	−0.327	0.104	−0.222
	Deepseek	0.000	0.138	0.000	0.053	0.191	−1.848	−0.691	−1.810	−0.750	−5.099
Medical	Llama	1.000	0.463	1.000	0.714	3.178	1.202	−0.404	1.031	0.340	2.169
	Deepseek	0.636	1.000	0.594	0.838	3.068	0.172	1.127	−0.111	0.691	1.879
	Gemini	0.848	0.854	0.943	0.419	3.064	0.773	0.710	0.870	−0.497	1.855
	Claude	0.000	0.707	0.630	1.000	2.338	−1.632	0.292	−0.009	1.150	−0.198
	OpenAI	0.394	0.000	0.000	0.000	0.394	−0.515	−1.726	−1.780	−1.684	−5.706

Table 2. This table provides a domain-wise performance comparison of five large language models across five domains using normalized evaluation methods. Both Min-Max scaling and Z-score normalization are applied to four core metrics semantic similarity (Sem), sentiment neutrality (Sent), TF-IDF similarity (TF-IDF), and named entity recognition accuracy (NER) to derive aggregate scores and ranks.

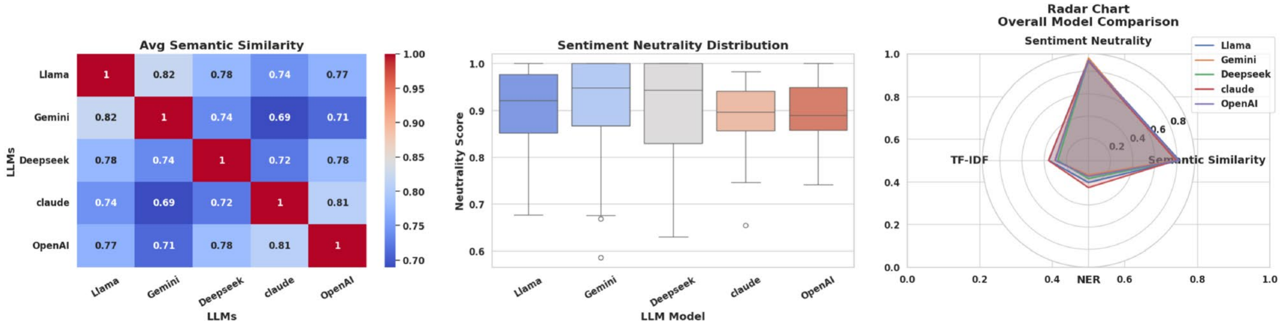

Fig. 4. This figure offers a comparative analysis of LLM performance. The left heatmap presents average semantic similarity scores across models, indicating how closely aligned their outputs are. The center box plot shows the distribution of sentiment neutrality scores, revealing the consistency of objective responses. The right radar chart summarizes performance across four key metrics semantic similarity, sentiment neutrality, TF-IDF, and NER accuracy providing a holistic view of each model's strengths and trade-offs.

Claude and OpenAI show moderate results, with some variation depending on the evaluation metric. DeepSeek consistently ranks last, underperforming in semantic similarity, factual grounding, and hallucination detection. The strong performance of Llama in semantic similarity, TF-IDF and NER score alignment underscores its ability to handle agricultural queries with precision and factual robustness.

Biology

When aggregating results across all queries in the biology domain as displayed in Figure 4 and Table 1, Llama again leads with the highest semantic similarity score (0.822), slightly ahead of OpenAI (0.814), and Gemini and Claude trail closely at 0.791. Sentiment analysis consistently shows high neutrality scores for all models, confirming scientific responses' expected neutrality. Claude achieves the best TF-IDF similarity (0.361) and NER-based factual accuracy (0.245), reflecting excellent alignment with the source material. Llama also performs consistently well in both semantic and factual evaluations, while DeepSeek remains at the lower end.

Overall, Llama and Claude proved to be the most reliable models for biology-related queries. According to Table 2, the final rankings for the biology domain place Llama at the top, securing first place in both min-max and z-score normalization evaluations. Claude follows closely behind, demonstrating strong, balanced performance across all metrics. OpenAI and Gemini fight between to secure ranks third and fourth, maintaining moderate and steady results. Meanwhile, DeepSeek consistently occupies the lower ranks in antecedent similarity, sentiment neutrality, and hallucination detection, indicating less reliable output. In conclusion, Llama and Claude emerge as the most trustworthy models for addressing biology-focused queries with both semantic accuracy and factual rigor.

Economics

Referring to Table 1 and Figure 5, the broader evaluation across economics-related queries reveals that Llama leads with a semantic similarity score of 0.761, trailed by Gemini at 0.728, while Claude registers the lowest score of 0.701. Sentiment analysis continues to show uniformly neutral outputs, as expected for technical and policy-focused content. In factual consistency metrics, Llama once again leads, achieving the highest TF-IDF similarity (0.426) and NER accuracy (0.205), reflecting strong grounding in the source material and reliable entity recognition. DeepSeek consistently underperforms across both factual verification metrics, indicating higher rates of hallucination and lower adherence to original content. Overall, Llama demonstrates the most balanced performance across both semantic and factual dimensions for economics-related queries.

Table 2 confirms Llama's dominance in the economics domain, where it ranks first using both min-max normalization and z-score normalization methods. Gemini claims second place with strong performance across most metrics, while Claude lands in third with stable but moderate results. OpenAI and DeepSeek occupy the lower positions across all evaluation measures. Llama's consistent strength in semantic similarity, TF-IDF-based alignment, and NER factual accuracy firmly establishes it as the most dependable model for addressing complex economic research queries.

IOT

When analyzing all IoT domain queries collectively, Llama emerges as the leading model, achieving the highest semantic similarity (0.837), TF-IDF similarity (0.444), and NER accuracy (0.501), as highlighted in Table 1 and Figure 6. OpenAI and Claude also perform well, with OpenAI ranking second in semantic similarity (0.832) and Gemini ranking second in TF-IDF similarity (0.432), while Claude demonstrates notable strength, particularly in NER accuracy (0.395). Gemini shows moderate performance, achieving a semantic similarity score of 0.822 and NER accuracy of 0.368, indicating solid, though not leading, results. DeepSeek consistently underperforms, especially in TF-IDF similarity (0.210), highlighting greater lexical hallucination. Sentiment neutrality remains low across all models, consistent with expectations for technical IoT-focused content. Overall, Llama stands out as the most reliable and factually consistent model for IoT queries, with Gemini and Claude providing strong secondary support. As illustrated in Table 2, the final rankings in the IoT domain reaffirm Llama's position at the top, securing first place based on both min-max and z-score normalization due to its consistently strong performance across semantic similarity, factual grounding, and bias neutrality. Gemini claims second place, thanks to solid semantic alignment and moderate factual reliability. Claude ranks third, performing well in NER-based evaluations but slightly trailing in semantic similarity compared to the leaders. OpenAI and DeepSeek occupy the lower ranks, showing weaker results across most metrics. In summary, Llama proves to be the most capable and balanced model for handling IoT-related queries among all the evaluated LLMs.

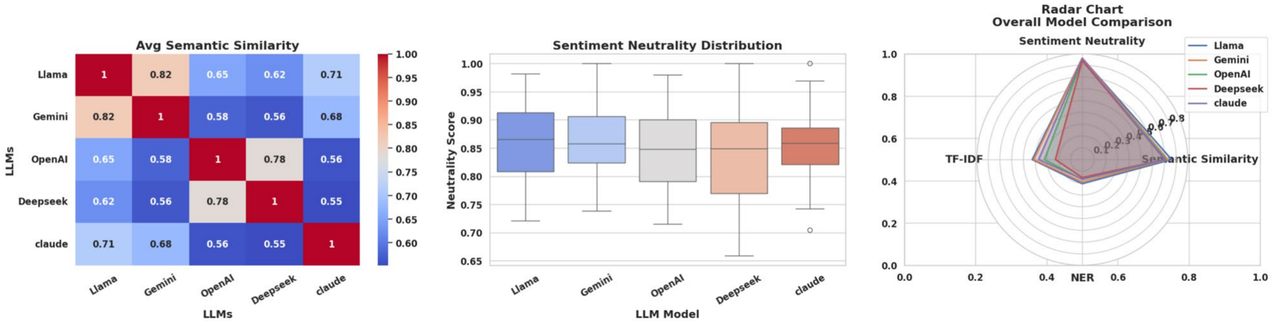

Fig. 5. This figure provides a comparative performance overview of multiple LLMs. The left heatmap illustrates the average semantic similarity between models, revealing how closely their responses align. The middle box plot displays the distribution of sentiment neutrality scores, highlighting each model's consistency in generating unbiased content. The right radar chart integrates key metrics semantic similarity, sentiment neutrality, TF-IDF, and NER accuracy into a single visual, offering an at-a-glance comparison of overall model performance.

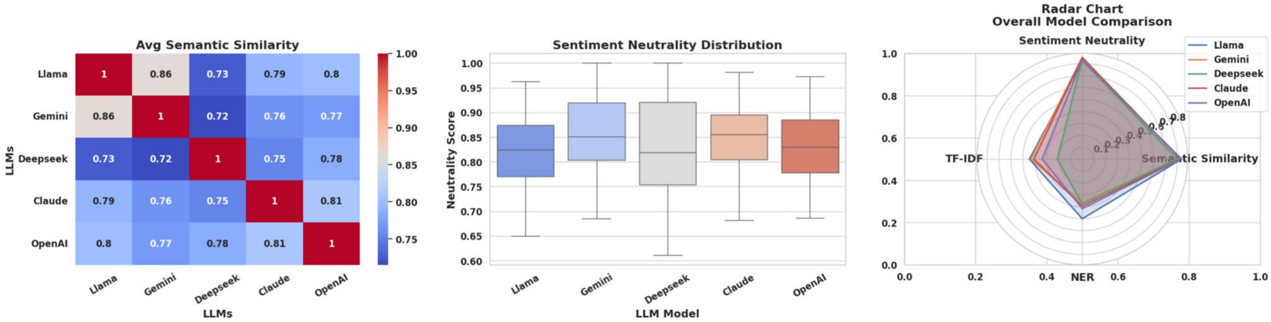

Fig. 6. This figure presents a comparative analysis of LLMs using multiple evaluation metrics. The left heatmap shows the average semantic similarity scores across models, reflecting how closely their responses align in meaning. The middle box plot displays sentiment neutrality distributions, indicating each model's ability to generate unbiased and objective content. The right radar chart offers an integrated view of model performance across four key metrics: semantic similarity, sentiment neutrality, TF-IDF similarity, and NER-based accuracy, facilitating holistic model comparison.

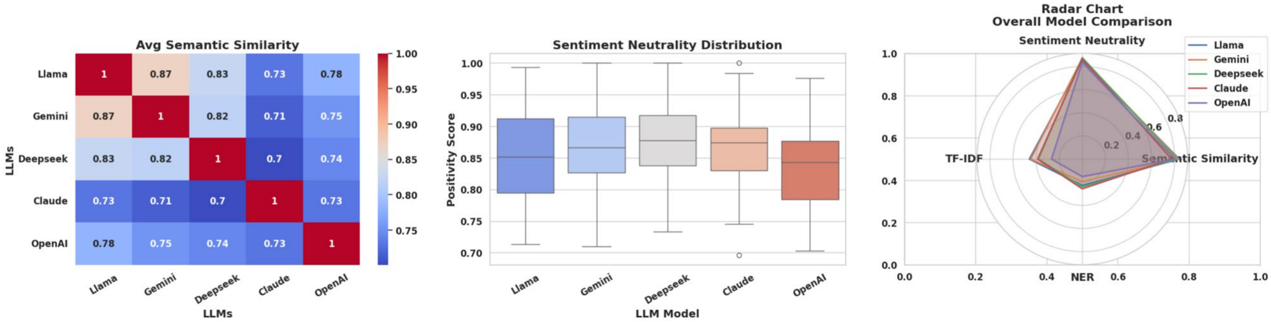

Fig. 7. This figure compares LLM performance using three visualizations. The left heatmap illustrates the average semantic similarity between models, indicating the alignment of their outputs in terms of meaning. The middle box plot shows the distribution of sentiment positivity scores, capturing how positively each model responds. The right radar chart provides an integrated performance view across semantic similarity, sentiment neutrality, TF-IDF similarity, and NER accuracy, enabling a comprehensive comparison of model strengths.

Medical

The broader evaluation of all medical domain queries is illustrated in Figure 7 and Table 1. Llama maintains the highest overall semantic similarity score (0.841), followed closely by Gemini (0.831). Claude records the lowest semantic similarity (0.775) among the evaluated models. Sentiment analysis continues to show uniformly neutral outputs, as expected for technical and policy-focused content. In factual consistency metrics, Llama once again leads, achieving the highest TF-IDF similarity (0.411), reflecting strong grounding in the source material. Overall, Llama demonstrates the most balanced performance across both semantic and factual dimensions for medical-related queries.

Table 2 confirms Llama's leadership in the final rankings for the medical domain, as it secures the top position under both min-max normalization and z-score normalization evaluation strategies. Deepseek claims second place with strong performance across most metrics, while Gemini lands in third with stable but moderate results. Claude and OpenAI occupy the lower positions across all evaluation measures. Llama's consistent strength in semantic similarity, TF-IDF-based alignment, and NER factual accuracy firmly establishes it as the most dependable model for addressing complex medical research queries.

Overall comparison

The major insights and findings reveal that our comprehensive evaluation of five leading large language models (LLMs) across diverse domains - agriculture, biology, economics, IoT, and medical - uncovers distinct performance patterns and demonstrates significant variations in model capabilities across different specializations. This assessment, grounded in metrics like semantic similarity, sentiment neutrality, TF-IDF similarity (reflecting factual grounding), and NER-based accuracy (capturing entity recognition), offers a comprehensive, data-driven perspective on the capabilities and shortcomings of Llama, Gemini, Claude, OpenAI, and DeepSeek. Figure 8 illustrates the Semantic and NER Score heatmap across all domains. According to Figure 9 and Table 3 Llama emerges as the standout model, achieving the highest average final score of 1.629. Its dominance is fueled by leading scores in semantic similarity (0.786), TF-IDF alignment (0.878), and NER accuracy (0.416), highlighting

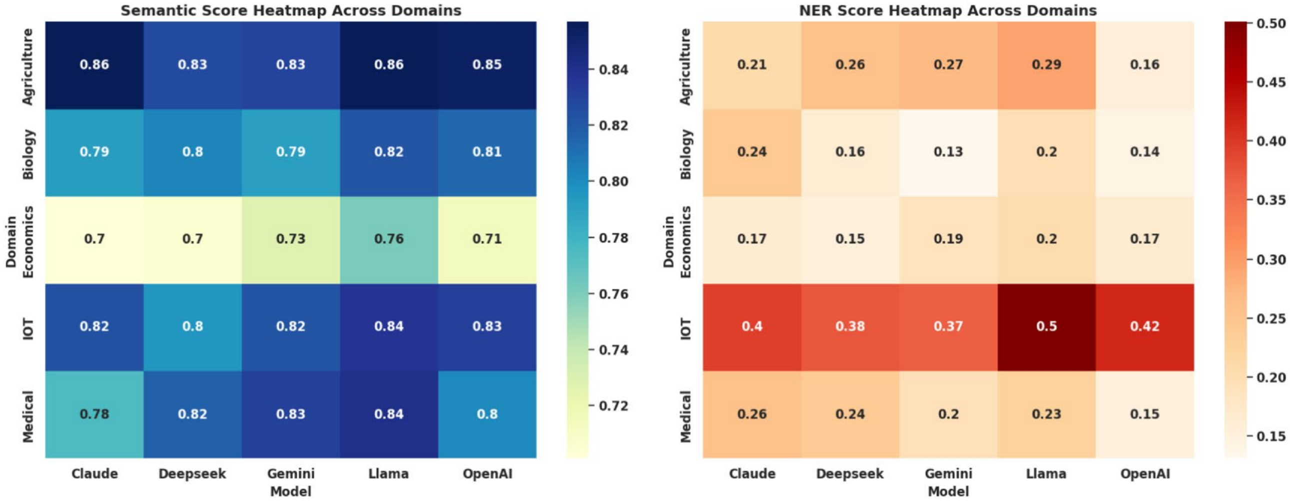

Fig. 8. Comparative Heatmap of Semantic and Named Entity Recognition Scores Illustrating Domain-Specific Strengths of Five Language Models.

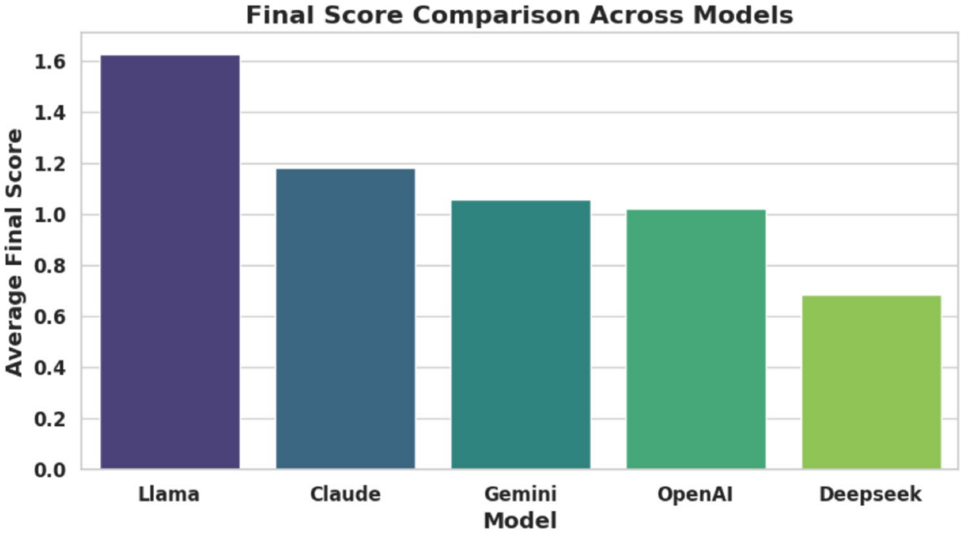

Fig. 9. Model-Wise Aggregated Score Visualization Reflecting General Effectiveness and Robustness Across Evaluation Metrics.

Model	Semantic	Sentiment	TF-IDF	NER	Final Score
Llama	0.786	0.451	0.878	0.416	1.629
Claude	0.569	0.460	0.740	0.334	1.183
Gemini	0.637	0.617	0.770	0.270	1.060
OpenAI	0.653	0.330	0.494	0.206	1.023
Deepseek	0.568	0.443	0.271	0.289	0.686

Table 3. Average performance of five language models across Semantic, Sentiment, TF-IDF, and NER tasks.

its strength in producing contextually accurate, factually grounded, and entity-rich responses. Though its sentiment neutrality score (0.451) is moderate, this neutrality is well-suited for technical and scientific discourse. Trailing Llama, Claude, and Gemini earn final scores of 1.183 and 1.060, respectively. Claude demonstrates balanced strength across all evaluation metrics, particularly excelling in factual coherence. Gemini, while scoring slightly lower in NER score (0.270), compensates with strong sentiment and TF-IDF results.

Domain	Model	Final Score
Agriculture	Llama	0.716
Biology	Llama	0.508
Economics	Llama	0.531
IOT	Llama	0.957
Medical	Llama	0.671

Table 4. Best-performing model in each domain based on final score. Llama consistently leads across all domains, showing strong cross-domain effectiveness.

OpenAI and DeepSeek round out the rankings, with final scores of 1.023 and 0.686. Although both models show moderate performance in semantic similarity, they struggle in sentiment analysis, TF-IDF and NER-based metrics, indicating weaknesses in maintaining factual correctness and precise language, particularly critical in fields like healthcare and economics. A deeper domain-specific analysis, as detailed in Table 4, confirms Llama's versatility, with the model leading in agriculture (0.716), biology (0.508), economics (0.531), IoT (0.957), and medical (0.671) domains. Semantic similarity heatmaps further illustrate Llama's consistent excellence, particularly in agriculture (0.86), IoT (0.84), and medical (0.84). While Gemini and OpenAI show strong results in certain areas, neither matches Llama's across-the-board consistency.

Overall, these findings emphasize the necessity of using multi-metric evaluation frameworks when choosing LLMs for knowledge-intensive tasks. High semantic similarity ensures contextual precision, while strong TF-IDF and NER metrics safeguard factual reliability and domain-specific expertise-critical factors for deploying LLMs effectively across diverse fields such as agriculture, biology, economics, medical, and IoT.

A comparative analysis of five prominent LLMs Llama, Gemini, Claude, OpenAI's GPT-4 Turbo, and DeepSeek reveals clear performance variations. Llama, in particular, demonstrates strong and consistent performance across all examined domains, suggesting a high degree of adaptability and general-purpose capability. The findings also reveal that some models are designed as generalists, while others excel in specific fields, likely due to differences in training data composition and model architecture. Training data quality appears to be a major factor influencing model performance. Models like Llama and Gemini show high semantic coherence and relatively low rates of factual error, which can be attributed to well-curated and balanced training datasets. On the other hand, DeepSeek exhibits weaker performance on TF-IDF and NER metrics, which may stem from a reliance on broader, less domain-focused data. This can lead to more frequent factual inconsistencies, particularly in complex technical domains. Sentiment analysis further supports the idea that models trained on domain-specific content tend to generate more neutral and objective responses a desirable characteristic for academic and technical discourse.

Limitations

While the MultiLLM-Chatbot framework offers a structured way to evaluate LLMs, several limitations should be acknowledged. The dataset, which consists of 50 research articles across five domains, is balanced but may not fully capture the breadth of scholarly writing, limiting how broadly our findings can be applied. Additionally, the 1,250 model responses, while diverse, may still carry biases related to source geography, discipline, or annotation. Our hallucination detection approach, based on TF-IDF and NER alignment, effectively flags surface-level errors but may miss deeper issues like paraphrased misinformation or logical gaps, which is especially concerning in sensitive fields like medicine or law.

Conclusion

This study presents a comprehensive evaluation of five leading large language models (LLMs) Llama, Gemini, Claude, OpenAI, and DeepSeek across five critical domains: agriculture, biology, economics, IoT, and medical. Employing a unified framework that integrates semantic similarity, sentiment analysis, TF-IDF similarity, and named entity recognition (NER) for factual accuracy, the analysis offers a nuanced understanding of each model's strengths and weaknesses within specific domain contexts. The findings consistently highlight Llama as the most robust and adaptable model, achieving top average scores across most domains. Its strong performance in semantic coherence, factual grounding, and entity recognition underscores its capability to generate contextually accurate and reliable outputs, making it highly suitable for a wide range of knowledge-intensive tasks. Gemini and Claude also perform competitively. In contrast, OpenAI and Deepseek exhibit moderate semantic capabilities but struggle with factual accuracy. The domain-wise analysis clearly shows that Llama stands out as the top performer across all fields, showcasing its strong ability to handle diverse language and content challenges. On the other hand, models like Claude, Gemini, and OpenAI show mixed results; they perform well in some domains but fall short in others. This highlights the need for choosing models not just based on overall performance but also on how well they handle specific domain requirements. Additionally, the findings emphasize the value of using a multi-dimensional evaluation approach. Focusing only on semantic similarity doesn't give the full picture, as it misses key aspects like factual accuracy, language variation, and proper entity recognition. By combining several evaluation metrics semantic, lexical, and entity-based we get a clearer, more reliable view of how each model performs in practical, real-world use cases.

As part of our future work, we plan to expand the scope of our evaluation by incorporating more diverse large language models and enriching our dataset with multilingual and culturally varied content. This will help reduce the English-language bias currently observed in most LLMs and improve generalization across languages and domains. We also aim to enhance our query generation process with expert input and refine hallucination detection using advanced techniques such as claim verification and human-in-the-loop validation. These improvements will make our framework more robust, inclusive, and better suited for real-world applications.

Data availability

Some or all data, model, or codes that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 9 May 2025; Accepted: 6 August 2025
Published online: 16 August 2025

References

1. Abdulla, H., Eltahir, A.M., Alwahaishi, S., Saghair, K., Platos, J. & Snasel, V. Chatbots development using natural language processing: a review. In: 2022 26th International Conference on Circuits, Systems, Communications and Computers (CSCC), pp. 122–128 (2022). IEEE
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. Attention is all you need. Advances in neural information processing systems **30** (2017)
3. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. & Mian, A. A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435 (2023)
4. OpenAI: GPT-4 Turbo Technical Report. OpenAI. https://platform.openai.com/docs/models/gpt-4-turbo (2024)
5. Anthropic: Claude 3.7 Sonnet Model Card. Anthropic. https://www.anthropic.com/claude/sonnet (2024)
6. DeepMind, G. Gemini 2.0 Flash Model Overview. Google DeepMind. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash (2024)
7. AI, M. LLaMA 3.3-70B Versatile Model Card. Meta AI. https://console.groq.com/docs/model/llama-3.3-70b-versatile (2024)
8. AI, D. DeepSeek R1 Zero Technical Report. DeepSeek AI. https://api-docs.deepseek.com/news/news250120 (2024)
9. Xu, S., Wu, Z., Zhao, H., Shu, P., Liu, Z., Liao, W., Li, S., Sikora, A., Liu, T. & Li, X. Reasoning before comparison: Llm-enhanced semantic similarity metrics for domain specialized text analysis. arXiv preprint arXiv:2402.11398 (2024)
10. Lin, L., Wang, L., Guo, J. & Wong, K.-F. Investigating bias in llm-based bias detection: Disparities between llms and human perception. arXiv preprint arXiv:2403.14896 (2024)
11. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**(8017), 625–630 (2024).
12. Chen, J., Lin, H., Han, X. & Sun, L. Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 17754–17762 (2024)
13. Bommasani, R., Liang, P. & Lee, T. Holistic evaluation of language models. *Annals of the New York Academy of Sciences* **1525**(1), 140–146 (2023).
14. Lin, S., Hilton, J. & Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958 (2021)
15. Guu, K., Lee, K., Tung, Z., Pasupat, P. & Chang, M. Retrieval augmented language model pre-training. In: International Conference on Machine Learning, pp. 3929–3938 (2020). PMLR
16. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **33**, 9459–9474 (2020).
17. Khattab, O. & Zaharia, M. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 39–48 (2020)
18. Nogueira, R. & Cho, K. Passage re-ranking with bert. arXiv preprint arXiv:1901.04085 (2019)
19. Reimers, N. & Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
20. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R. & Deng, L. Ms marco: A human-generated machine reading comprehension dataset (2016)
21. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. & Gurevych, I. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. arXiv preprint arXiv:2104.08663 (2021)
22. Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V. & Kalai, A.T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems **29** (2016)
23. Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P.M. & Bowman, S.R. Bbq: A hand-built bias benchmark for question answering. arXiv preprint arXiv:2110.08193 (2021)
24. Hutto, C. & Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, pp. 216–225 (2014)
25. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)
26. Awlla, K.M., Veisi, H. & Abdullah, A.A. Sentiment analysis in low-resource contexts: Bertâ€™s impact on central kurdish. Language Resources and Evaluation, 1–31 (2025)
27. Torge, S., Politov, A., Lehmann, C., Saffar, B. & Tao, Z. Named entity recognition for low-resource languages-profiting from language families. In: Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), pp. 1–10 (2023)
28. Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A. & Matias, Y. True: Re-evaluating factual consistency evaluation. arXiv preprint arXiv:2204.04991 (2022)
29. Kryściński, W., McCann, B., Xiong, C. & Socher, R. Evaluating the factual consistency of abstractive text summarization. arXiv preprint arXiv:1910.12840 (2019)
30. Keraghel, I., Morbieu, S. & Nadif, M. A survey on recent advances in named entity recognition. arXiv preprint arXiv:2401.10825 (2024)
31. Gupta, V., Pantoja, D., Ross, C., Williams, A. & Ung, M. Changing answer order can decrease mmlu accuracy. arXiv preprint arXiv:2406.19470 (2024)
32. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615 (2022)
33. Chakraborty, S., Chowdhury, R., Shuvo, S.R., Chatterjee, R. & Roy, S. MultiLLM-Chatbot: A Scalable Framework for Cross-Domain RAG-Based Evaluation and Hallucination Detection Using Multiple LLMs. https://drive.google.com/file/d/1_eK2XwODWGYsbzxevHn8UO-dP2CE1o_R/view. Accessed: 2025-02-20. Our observations indicate that the evaluated models are primarily trained in English and show reduced performance on less-resourced or non-English languages. (2025)

34. Okhapkin, V.P., Okhapkina, E.P., Iskhakova, A.O. & Iskhakov, A.Y. Constructing of semantically dependent patterns based on spacy and stanfordnlp libraries. In: Futuristic Trends in Network and Communication Technologies: Third International Conference, FTNCT 2020, Taganrog, Russia, October 14–16, 2020, Revised Selected Papers, Part I 3, pp. 500–512 (2021). Springer
35. Cohan, A., Beltagy, I., King, D., Dalvi, B. & Weld, D.S. Pretrained language models for sequential sentence classification. arXiv preprint arXiv:1909.04054 (2019)
36. Ni, C., Wu, J., Wang, H., Lu, W. & Zhang, C. Enhancing cloud-based large language model processing with elasticsearch and transformer models. In: International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2024), vol. 13180, pp. 1648–1654 (2024). SPIE
37. Malkov, Y. A. & Yashunin, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* **42**(4), 824–836 (2018).
38. Khattab, O., Santhanam, K., Li, X.L., Hall, D., Liang, P., Potts, C. & Zaharia, M. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. arXiv preprint arXiv:2212.14024 (2022)
39. Chakraborty, S., Chowdhury, R., Shuvo, S.R., Chatterjee, R. & Roy, S. Sample Queries and Responses from LLMs. https://drive.google.com/file/d/1FEg1mJcmT699_-IxojoNpm7ZbLtGb-eV/view. Additional supplementary material demonstrating query-response pairs across evaluated domains. (2025)

Author contributions

Sorup Chakraborty, Rajesh Chowdhury, Sourov Roy Shuvo, and Rajdeep Chatterjee contributed equally to the conceptualization, methodology design, and data analysis. Rajdeep Chatterjee and Satyabrata Roy supervised the research, provided critical revisions, and finalized the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding

Open access funding provided by Manipal University Jaipur. No funding was received for conducting this study.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.