International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-14, Issue-5, Sep-Oct 2026
DOI: https://doi.org/10.55524/ijircst.2026.14.5.3
Article ID IRP-1787, Pages 19-33
www.ijircst.org

Evaluating the Efficiency of Large Language Models in Detecting and Mitigating Hallucinations Using a Curated Domain-Oriented Dataset and Post-Hoc Retrieval

Chandra Sekhar Sanaboina[1], and Tammisetti Surekha[2]

[1]Assistant Professor, Department of Computer Science and Engineering, UCEK, JNTUK, Kakinada, Andhra Pradesh, India
[2]PG Scholar, Department of Computer Science and Engineering, UCEK, JNTUK, Kakinada, Andhra Pradesh, India

Correspondence should be addressed to Chandra Sekhar Sanaboina; chandrasekhar.s@jntucek.ac.in

Received: 25 August 2026; **Revised:** 8 Sep 2026; **Accepted:** 19 Sep 2026; **Published Online:** 24 Sep 2026

ABSTRACT- Hallucination in large language models (LLMs) refers to the generation of plausible and coherent information that is incorrect, unsupported, or inconsistent with the available evidence. Detecting such hallucinations and verifying the information in generated responses are important for improving the reliability of LLM based systems. This study evaluates the ability of LLMs to detect hallucinated responses and investigates whether external evidence can improve the verification process. A curated dataset of 700 question and answer pairs covering 20 academic disciplines was used, including both factual and hallucinated responses. Four models, GPT 3.5 Turbo, GPT 4, GPT 4 Turbo, and Gemma 7B, were first evaluated using their internal knowledge. For responses that required further checking, the responses were divided into individual claims, and relevant evidence was retrieved from Wikipedia using BM25 and semantic re ranking. Each claim was then verified against the retrieved evidence, and FActScore was used to support the final classification as factual or hallucinated. With no evidence the models scored accuracies between 0.653 and 0.943. After adding evidence from Wikipedia accuracy rose from 0.88 to 0.986. Fleiss' Kappa also increased from 0.544 to 0.829, indicating greater agreement among the models after external evidence was introduced. These results show that relevant external evidence can improve the reliability and consistency of LLM based hallucination detection. The proposed approach uses retrieval to verify claims and reclassify responses when the initial judgment needs evidence. Retrieval does not directly correct the content. This checking step can therefore help mitigating hallucination by spotting information that is not supported.

KEYWORDS: Large Language Models, Hallucination Detection, Claim Verification, Evidence Retrieval, Hallucination Mitigation, FActScore, Semantic re-ranking, Factuality Assessment

I. INTRODUCTION

Large language models (LLMs) have changed how people interact with information systems. Large language models (LLMs) can understand language and produce detailed replies. That means Large language models (LLMs) can answer questions from fields, such, as education, healthcare, finance, legal services and scientific research. Yet Large language models (LLMs) can also create hallucinations. A hallucination is when a Large language model (LLM) gives information that sounds convincing but is actually wrong made up or not backed by evidence [12] [17]. This makes a problem because a wrong answer can look like a right one and it is hard for users to see the mistake without extra checking. The problem is worse where wrong information can spread misinformation or change decisions that rely on facts [1] [16].

This problem becomes more serious as LLMs increasingly evaluate the outputs of other LLMs under what is commonly called the LLM-as-a-Judge paradigm [20]. If the judge model is itself prone to hallucination, it may wrongly approve an unsupported claim or reject a response that was actually correct, which undermines the reliability of the entire evaluation pipeline [20]. Benchmarks such as HaluEval [3] and TruthfulQA [7] have already shown that even widely used models fabricate or repeat false information at a meaningful rate. To address this, retrieval-augmented verification has emerged as an effective approach, in which a model's judgment is grounded in external evidence rather than internal knowledge alone. Instead of relying solely on what the model already knows, the system retrieves relevant passages from a trusted source and checks each claim in the response against that evidence, an approach that has been shown to reduce hallucination and improve factual precision in prior work [4], [6], [11], [18].

In this work, we propose a retrieval-augmented, claim-level verification framework and evaluate it across four LLMs, namely GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, and Gemma 7B. Each response under evaluation is decomposed into atomic factual claims, following the FActScore methodology [6], so that every statement can be verified independently rather than judged as part of an undifferentiated block of text. For each claim, the corresponding evidence is retrieved from Wikipedia using BM25 lexical retrieval and refined through semantic re-ranking, allowing the system to combine keyword-level and meaning-level relevance when selecting supporting passages. This combination of claim decomposition and

two-stage retrieval improves the reliability of the resulting factual judgment and reduces the system's dependence on any single model's internal, and potentially outdated, knowledge.

The proposed framework allows a generated response to be classified as factual or hallucinated based on the proportion of its claims that are actually supported by retrieved evidence, rather than on the judge model's unaided confidence. By grounding classification in retrieved content rather than internal recall, the framework improves reliability and reduces the risk of a model confidently endorsing an unsupported claim. The evaluation dataset itself, curated from the MMLU benchmark across 20 academic disciplines with controlled hallucinations introduced into a portion of the responses, is designed to reflect the kind of domain diversity a hallucination detector would need to handle in practice [1], [2], [8]. This approach is intended to be model-agnostic, so it can be adapted across different LLM providers and extended to additional domains without changes to its underlying design.

II. LITERATURE REVIEW

Work on making LLM outputs more trustworthy has grown quickly, and one useful entry point is the structured question-answering setup with post-hoc retrieval presented in [1]. The study combines hallucination detection with semantic similarity analysis and cross-referencing of generated answers against source information, followed by retrieval of external evidence to correct detected hallucinations. Moreover, the study shows improvements in accuracy, precision, recall and F1-score with the use of post-hoc retrieval. However, [1] does not specify enough information on how the hallucination score was obtained, nor the threshold used to decide whether a response is hallucinated. This makes the detection stage less explicit and leaves scope for a more clearly defined and reproducible claim-level hallucination detection approach. Around the same idea, several efforts aimed at building larger and more careful hallucination benchmarks [2], [3], [7] make the case that detection research is only as good as the data behind it, and each push toward datasets that expose unsupported or fabricated content across medical, legal, financial, and general-knowledge questions rather than a single narrow domain.

Several survey papers [5], [12], [13], [17] step back from any one method and instead map out the field as a whole, sorting detection techniques into white-box, black-box, and external-knowledge categories and tracking how factuality evaluation has changed as models have gotten larger. A recurring theme across these surveys is that no single detection strategy is complete on its own, and that inconsistent metrics and the overlap between hallucination and bias make head-to-head comparison difficult. That concern is echoed by work looking specifically at reliability and adversarial robustness [9], [16], which shows that responses grounded in nothing but a model's own confidence remain vulnerable, particularly when fabricated details are deliberately slipped into a clinical or decision-support context.

A separate line of work checks models against themselves rather than against outside sources. Self-consistency and black-box detection techniques [4], [19] rely on the idea that a model's own repeated answers will agree when it actually knows something and drift apart when it does not, and both show this can catch hallucination without needing access to model internals. Even so, neither claims to match what grounding in real external evidence achieves, which is part of why they tend to be framed as a first line of defense rather than a complete solution.

On the measurement side, two developments stand out. A scoring method that breaks a response down into atomic, individually checkable claims [6] gives a far more precise picture of where a model goes wrong than a single correct-or-incorrect label ever could, and an automated approach to generating model-specific hallucination data [8] removes much of the manual annotation cost that has historically slowed this kind of research down. Between the two, both how hallucination is measured and how the data needed to study it gets built have moved forward considerably.

Elsewhere, work that pulls in structured external knowledge shows similar promise outside pure text generation. A fact-checking system built around a knowledge graph [11] finds that combining structured relational data with unstructured passages beats relying on either source by itself, and a more applied example from automated software test generation [10] shows that constraining an LLM's reasoning with structured references cuts down on hallucination even in a task that is not, on the surface, about facts at all.

A few more recent papers focus squarely on making retrieval-based verification itself work better. One reframes prompt construction as something to be optimized around the retrieved evidence rather than fixed in advance [18], and two summarization-focused approaches [14], [15] show that grounding an abstractive summary in extractive evidence, or checking it against generated question-answer pairs, can cut hallucination without hurting fluency. The common thread across all three is that retrieval only helps as much as the step that uses it is designed to make use of it.

One study looking directly at the LLM-as-a-Judge setup [20] found that fine-tuned judge models do well within the narrow domain they were trained on but struggle once asked to judge something unfamiliar, while stronger general-purpose models hold up more evenly across settings. That gap between working well in a familiar domain and actually generalizing is central to the motivation behind this paper, since it means a judge model's verdict cannot simply be trusted by default and needs to be checked against something outside itself.

This indicates that while LLMs offer strong capabilities as generators and evaluators, their tendency to hallucinate necessitates the integration of retrieval mechanisms, claim-level verification, and careful threshold-based evaluation rather than reliance on internal knowledge alone. This motivates the development of the retrieval-augmented, claim-level verification framework proposed in this paper, which draws on evidence retrieved from external sources to ensure accuracy, reliability, and consistency across models of differing strength. Furthermore, such a framework enables more dependable hallucination detection across diverse domains and question types by grounding every claim in externally checkable evidence rather than a single model's self-reported confidence. This also opens avenues for further improvement through adaptive, context-sensitive thresholds and evidence sources tuned to specific domains.

III. METHODOLOGY

The proposed framework follows an eight-stage pipeline that begins with the curated dataset, runs each response through an initial LLM-based judgment, and escalates to retrieval-based verification only when that judgment needs further checking. Full claim extraction, evidence retrieval and re-ranking happen for responses that the first check misclassifies. This avoids work on responses already judged correctly. Because the same target large language model is used for the judgment for claim verification and for generating Wikipedia titles the pipeline can work with any large language model. It can be swapped from one model to another without changing its structure. Figure 1 shows the architecture.

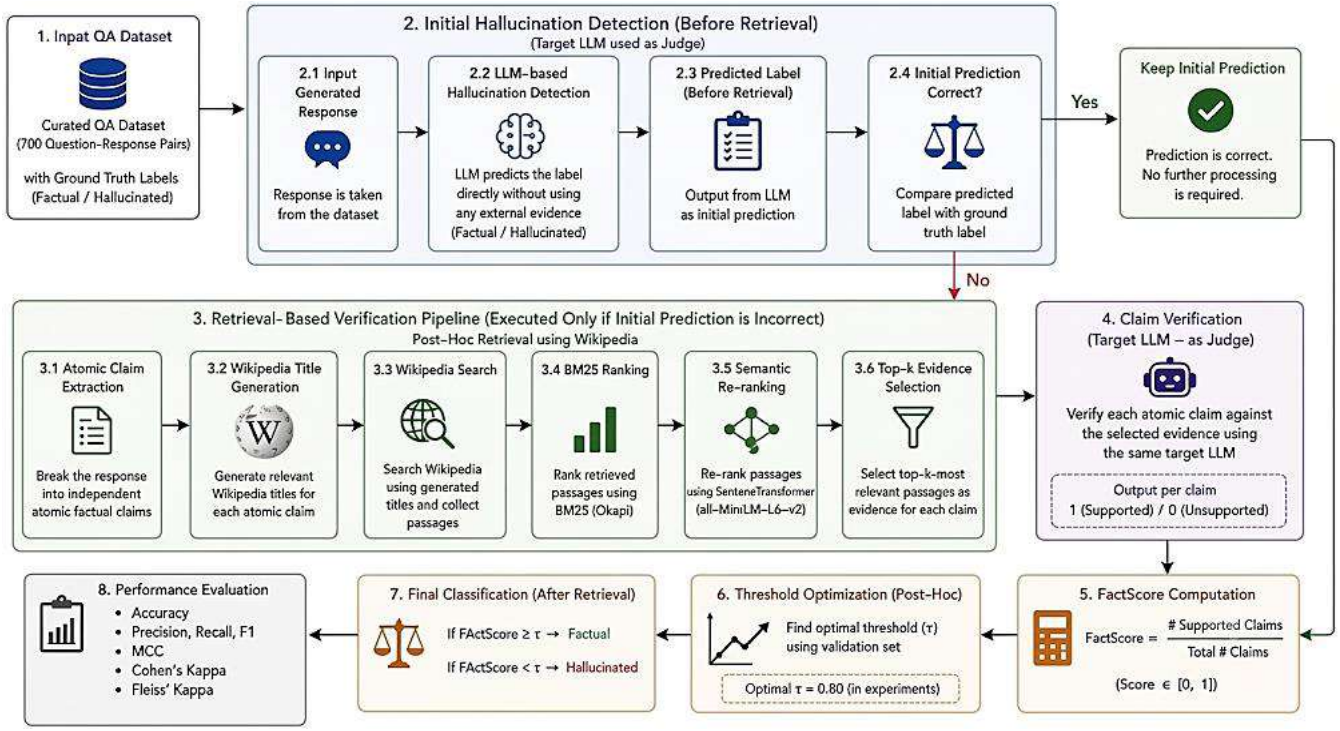


Figure 1: Architecture of the proposed retrieval-augmented claim-level hallucination-detection framework

A. *System Architecture*

Each response is first judged by the target LLM using only its internal knowledge; if that judgment already matches the ground truth, it is accepted as final. If not, the response is sent to a retrieval-based pipeline. This pipeline splits the response into claims. For each claim it generates a Wikipedia title. Then it. Ranks supporting passages using BM25 and semantic re-ranking. Each claim is verified against the retrieved evidence. The final FActScore is compared to a threshold, to the model to determine the outcome. Performance is measured using Accuracy, Precision, Recall, F1, MCC and Cohens Kappa.

B. *Dataset Preparation*

The dataset contains 700 question-response pairs built from questions drawn from the Massive Multitask Language Understanding (MMLU) benchmark, spanning 20 academic disciplines including Law, Computer Science, Artificial Intelligence, Data Science, Machine Learning, Mathematics, Physics, Chemistry, Biology, Medicine, Economics, Finance, Business Management, History, Geography, Political Science, Psychology, Philosophy, Environmental Science, and Literature. Each record pairs a question with a ground-truth answer and is labeled Factual or Hallucinated, with the two classes kept close to balanced. The hallucinated portion was created through four kinds of controlled perturbation: paraphrasing that subtly shifts meaning, insertion of contradictory statements, manipulation of a critical fact such as a date, name, or number, and mismatching of entities within an otherwise plausible answer.

C. *Initial LLM-as-a-Judge Evaluation*

Each generated response is first passed to the target model acting purely as a judge, using nothing but its own internal knowledge. Four models were evaluated: GPT-3.5 Turbo, GPT-4, GPT-4 Turbo and Gemma 7B. When a model's binary prediction matches the ground-truth label the framework accepts the model's prediction, as final. Otherwise, the sample is escalated to the retrieval-based pipeline. This approach focuses the retrieval budget, on cases where internal knowledge alone was not enough.

D. *Claim-Level Retrieval-Augmented Verification*

The retrieval pipeline first decomposes a response into independent atomic factual claims following the decomposition approach used in FActScore. For each of these claims the target language model comes up with the matching Wikipedia article title. That title is then used to find the sections from the article. These passages go through two filtering steps. The first step uses BM25 ranking, which looks for words that match well using the Okapi weighting scheme. This helps find sections, with lexical overlap. Next a semantic re-ranking is performed with Sentence-Transformer (all-MiniLM-L6-v2) embeddings, measuring cosine similarity. The top-k passages become the evidence set for the claim and the same target LLM acts as a claim-level judge returning 1 if the evidence supports the

claim and 0 otherwise. By repeating this process for every claim, the framework can pinpoint exactly which claims lead to hallucinated responses.

E. FActScore Computation and Threshold-Based Classification

Once all claims in a response have been verified, its FActScore is calculated as the ratio of the number of supported claims to the total number of claims, resulting in a score between 0 and 1. The FActScore is compared to a threshold to decide the classification. If the score is equal to or higher than the threshold the response is labeled as Factual. If the score is lower, than the threshold the response is labeled as Hallucinated. Various FActScore thresholds ranging from 0.30 to 0.95 were tested for each model. This helps find the optimal point to classify responses accurately. The threshold selection was based on the misclassified responses observed in the dataset. For each threshold, the changes in classification were examined along with the Precision and Recall of the Hallucinated class. The threshold was selected at the point where Precision and Recall provided a suitable balance while resulting in fewer classification errors. Since the FActScore distribution and misclassification patterns differed across the models, the selected threshold also differed for each model.

F. Performance Measures

The system is checked using seven ways: Accuracy, Precision, Recall, F1-score, MCC, Cohen's Kappa and Fleiss' Kappa. These ways together look at how correct the predictions how well they match the labels and how consistent the results are, across the models that are being tested.

For the six ways the Hallucinated class is treated as the class. True Positives (TP) are hallucinated answers that are correctly found. True Negatives (TN) are answers that are correctly found. False Positives (FP) are answers that are wrongfully found as hallucinated. False Negatives (FN) are hallucinated answers that are wrongfully found as factual.

Accuracy, which represents the proportion of correct predictions, is given by Equation (1):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision is the ratio of responses flagged as hallucinated that were actually hallucinated, while Recall is the ratio of truly hallucinated responses that were successfully identified, as shown in Equations (2) and (3):

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1-score is the harmonic mean of Precision and Recall, giving a single metric that balances both, punishing models that sacrifice one metric for the sake of the other. Equation (4) shows how the F1-score is calculated:

$$\text{F1 Score} = \frac{2\,(\text{Precision})\,(\text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)$$

The Matthews Correlation Coefficient (MCC) considers all four quantities of the confusion matrix at the same time and is informative in the presence of class imbalance. It returns a value between −1 and +1, where +1 means perfect prediction, 0 means random guessing, and −1 means total disagreement, as shown in Equation (5):

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

Cohen's Kappa measures agreement between predictions and ground truth after accounting for chance agreement, where p_o is observed agreement and p_e is expected agreement under random assignment, shown in Equation (6):

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)} \quad (6)$$

Fleiss' Kappa extends this idea to more than two raters, measuring the consistency of Factual/Hallucinated classifications across all four target LLMs simultaneously, shown in Equation (7):

$$\kappa_F = \frac{(\bar{P} - \bar{P}e)}{(1 - \bar{P}e)} \quad (7)$$

Here, $\bar{P}$ is the mean observed pairwise agreement across all four models, and $\bar{P}_e$ is the agreement expected under independent random classification. Equations (1)-(6) assess each model's individual reliability against ground truth, while Equation (7) assesses cross-model consistency independent of any single model's accuracy.

IV. RESULTS AND DISCUSSION

A. Model-Detected Predictions before Retrieval

This section shows the raw hallucination-detection output from each of the four models on the 700-instance dataset before any retrieval-based verification.

GPT-3.5 Turbo:

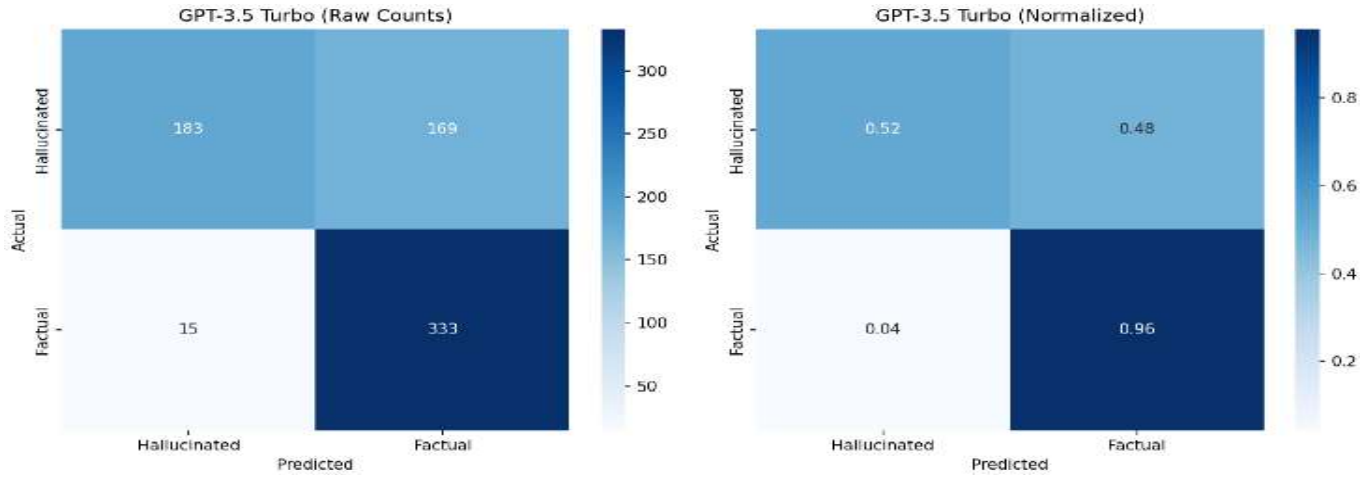


Figure 2: Confusion matrix of GPT-3.5 Turbo before retrieval

GPT-3.5 Turbo correctly detected 183/352 hallucinated responses (52%) and 333/348 factual responses (96%), as shown in Figure 2, indicating it was far more effective at recognizing factual than hallucinated responses before retrieval.

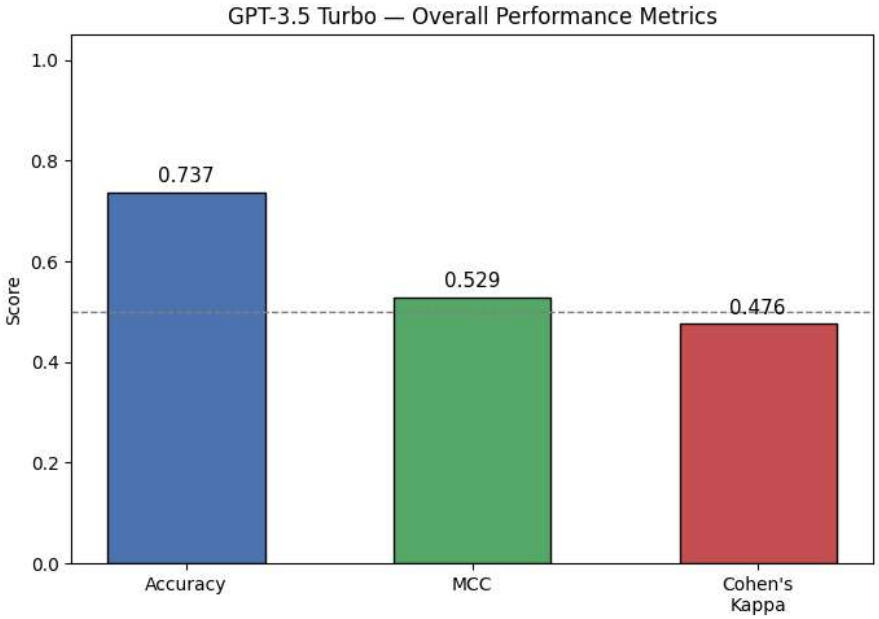


Figure 3: Overall performance metrics of GPT-3.5 Turbo before retrieval

As shown in Figure 3 the Accuracy, MCC and Cohen's Kappa scores for GPT-3.5 Turbo were 0.737, 0.529 and 0.476. These numbers represent a classification performance but the agreement with the ground truth was not perfect.

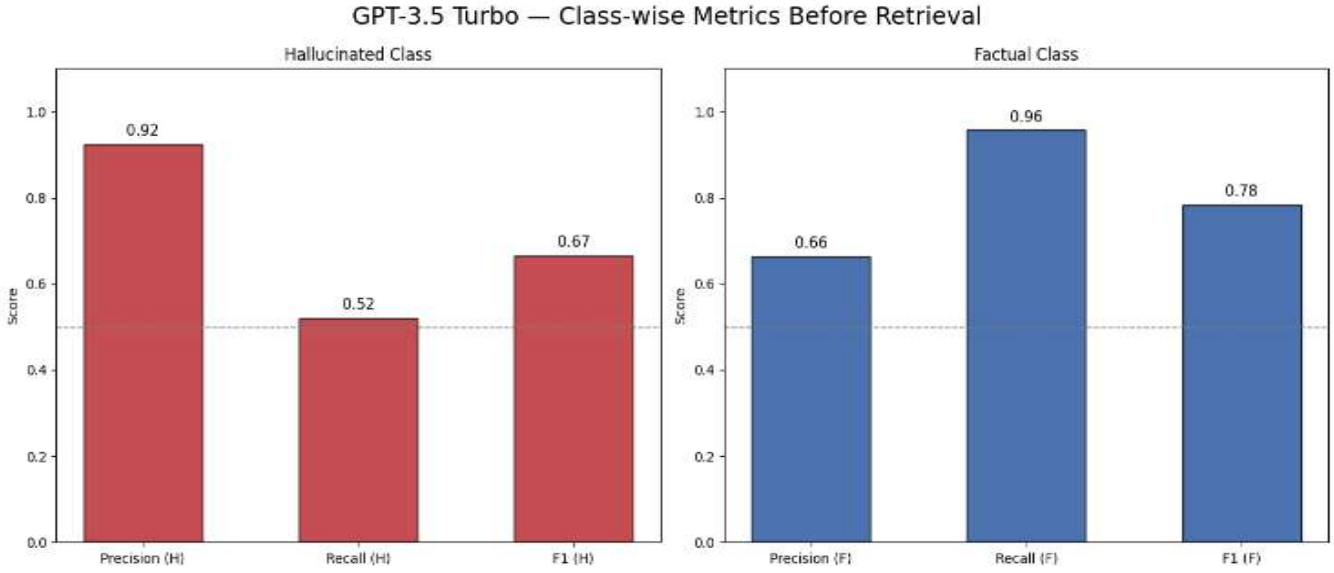


Figure 4: Class-wise metrics of GPT-3.5 Turbo before retrieval

GPT-3.5 Turbo achieved a precision of 0.92 recall of 0.52 and F1 score of 0.67, for the Hallucinated class. For the Factual class the precision was 0.66 recall was 0.96 and F1 score was 0.78 as shown in Figure 4. These results suggest that the model is good for identification of factual answers. However, a large number of hallucinated responses were not detected.

GPT-4:

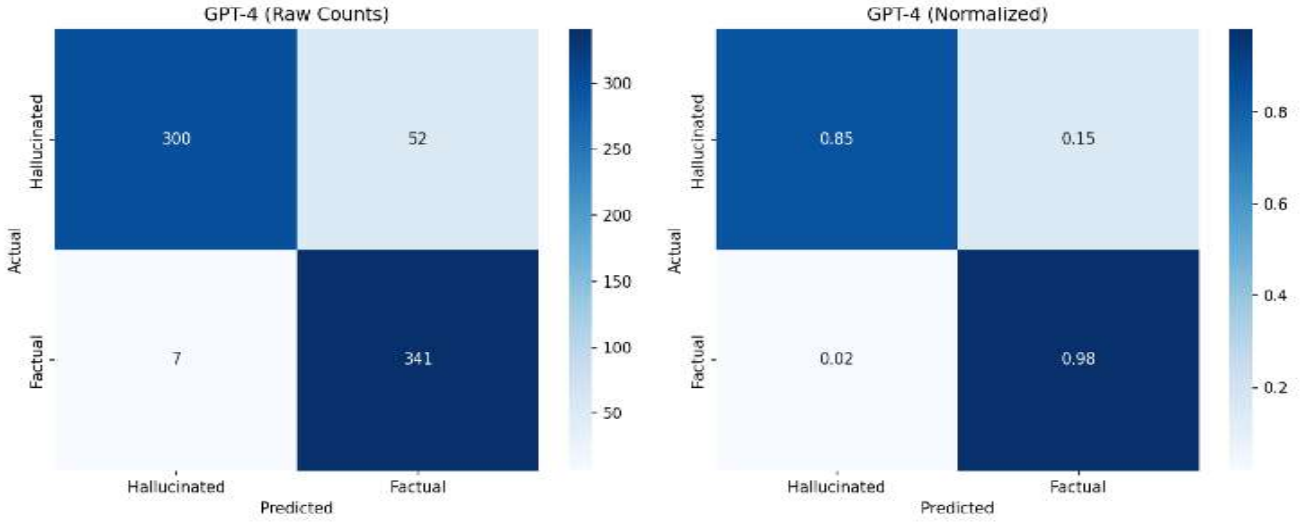


Figure 5: Confusion matrix of GPT-4 before retrieval

GPT-4 correctly identified 300/352 hallucinated (85%) and 341/348 factual (98%) responses, as shown in Figure 5, substantially stronger than GPT-3.5 Turbo at detecting hallucinations before retrieval.

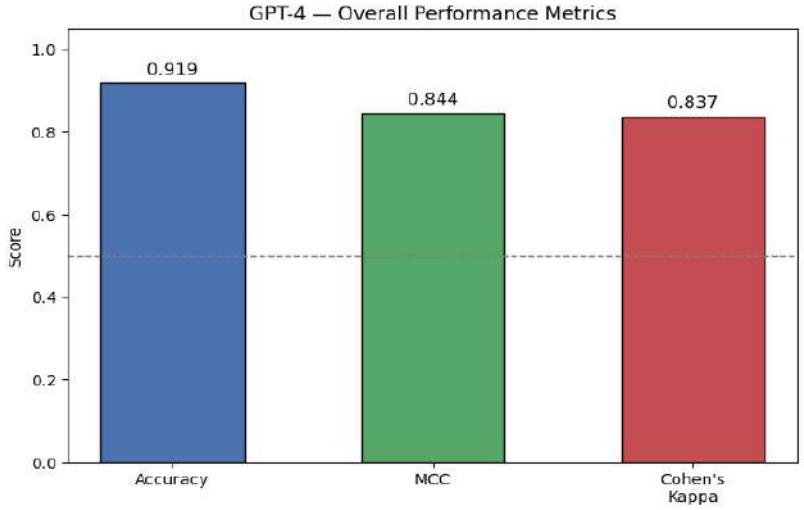


Figure 6: Overall performance of GPT-4 before retrieval

The Accuracy, MCC and Cohen's Kappa of GPT-4 were 0.919, 0.844 and 0.837 respectively, showing a high degree of agreement and a good baseline performance prior to retrieval, as shown in Figure 6.

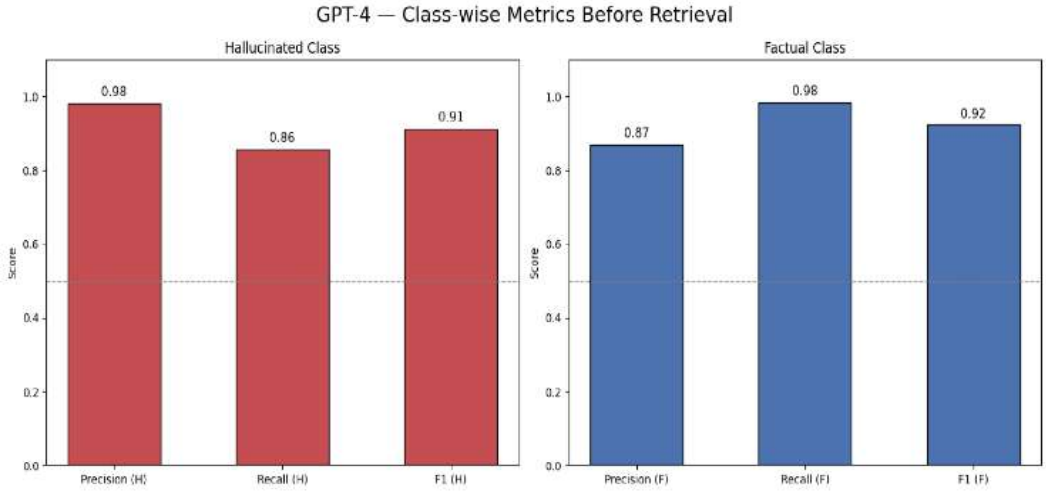


Figure 7: Class-wise metrics of GPT-4 before retrieval

GPT-4 achieved a precision of 0.98 recall of 0.86 and F1 score of 0.91 for the Hallucinated class. For the Factual class the precision was 0.87 recall was 0.98 and F1 score was 0.92 as shown in Figure 7. The results for both classes were similar with Precision, Recall and F1-score remaining relatively high.

GPT-4 Turbo:

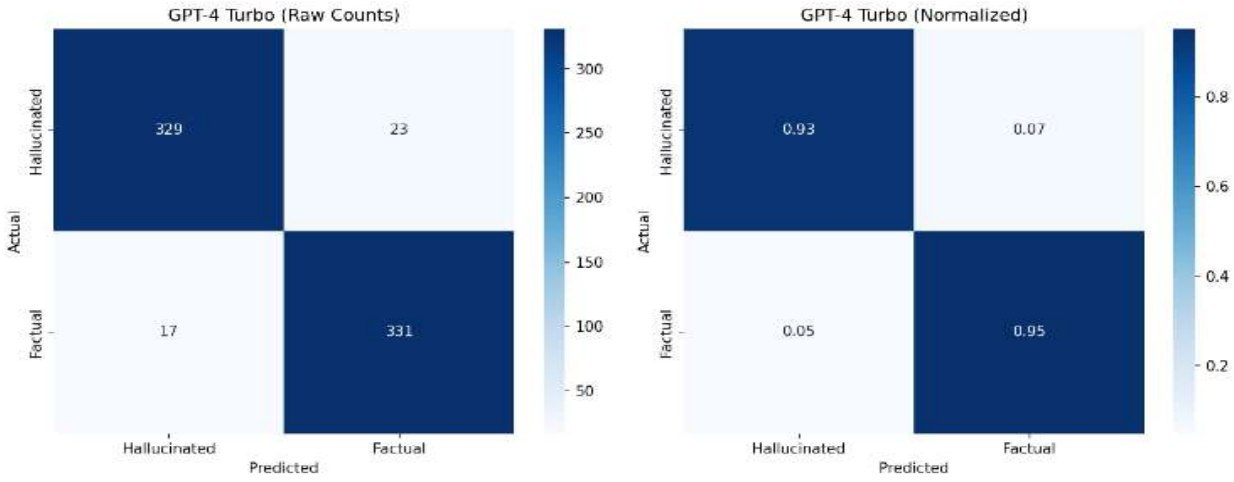


Figure 8: Confusion matrix of GPT-4 Turbo before retrieval

GPT-4 Turbo identified 329 of the 352 hallucinated responses correctly, corresponding to 93%, and correctly classified 331 of the 348 factual responses, corresponding to 95%, as shown in Figure 8. The results were similar for both classes, with only a small difference in the number of responses classified incorrectly.

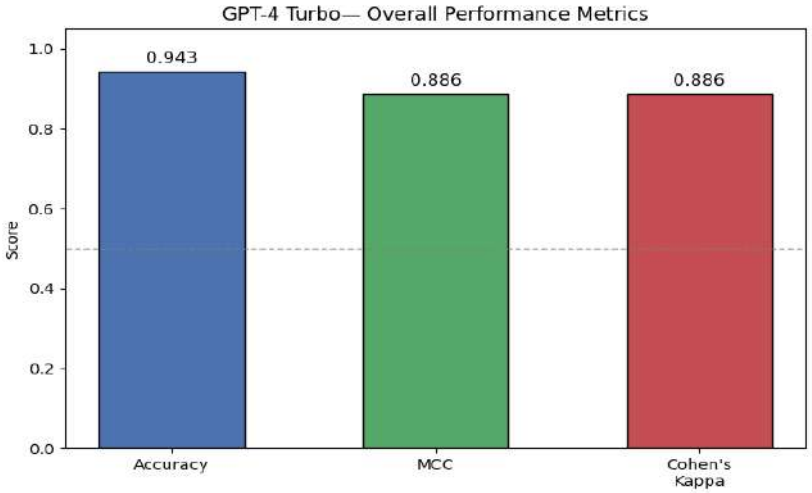


Figure 9: Overall performance of GPT-4 Turbo before retrieval

GPT-4 Turbo achieved an Accuracy of 0.943, an MCC of 0.886, and a Cohen's Kappa of 0.886, as shown in Figure 9. The values indicate that the model produced consistent results across the evaluated responses before retrieval.

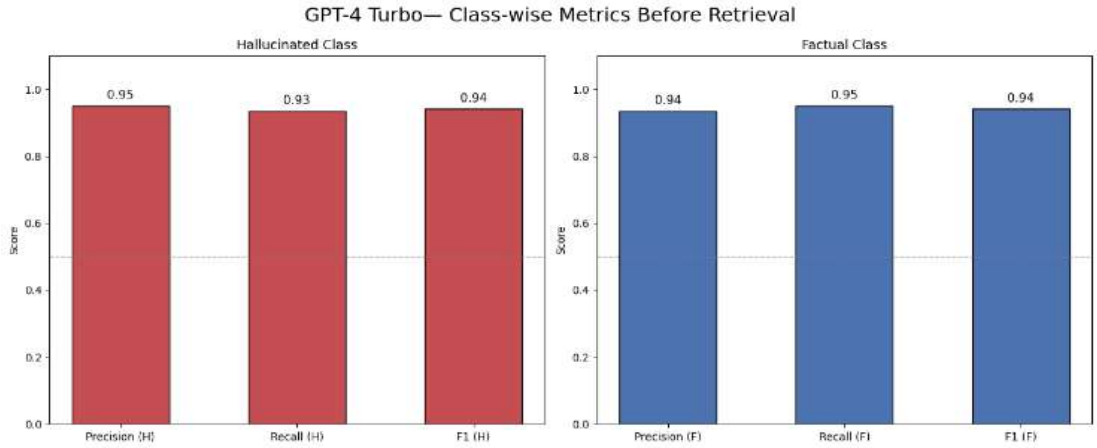


Figure 10: Class-wise metrics of GPT-4 Turbo before retrieval

GPT-4 Turbo achieved a precision of 0.95 recall of 0.93 and F1 score of 0.94 for the Hallucinated class. For the Factual class the precision was 0.94 recall was 0.95 and F1 score was 0.94 as shown in Figure 10. The results, for the two classes were very close each showing the F1-score of 0.94.

Gemma 7B:

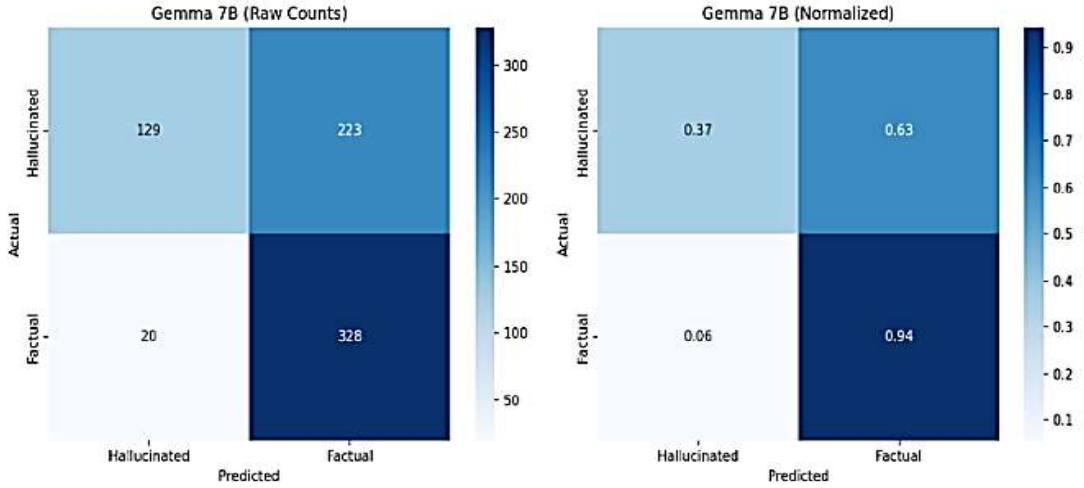


Figure 11: Confusion matrix of Gemma 7B before retrieval

Gemma 7B correctly identified 129/352 hallucinated (37%) and 328/348 factual (94%) responses, as shown in Figure 11, reflecting strong performance at recognizing factual responses but substantially weaker performance at detecting hallucinations before retrieval.

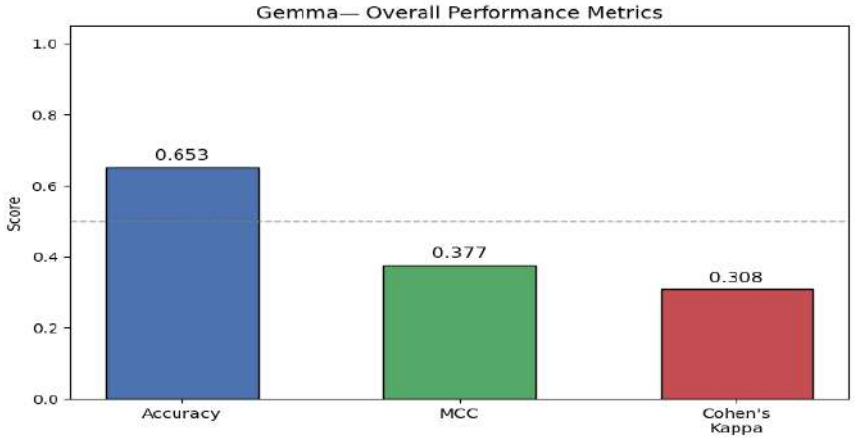


Figure 12: Overall performance metrics of Gemma 7B before retrieval

Gemma 7B achieved an Accuracy of 0.653, MCC of 0.377, and Cohen's Kappa of 0.308, as shown in Figure 12, notably lower than the other models, particularly in agreement with ground truth.

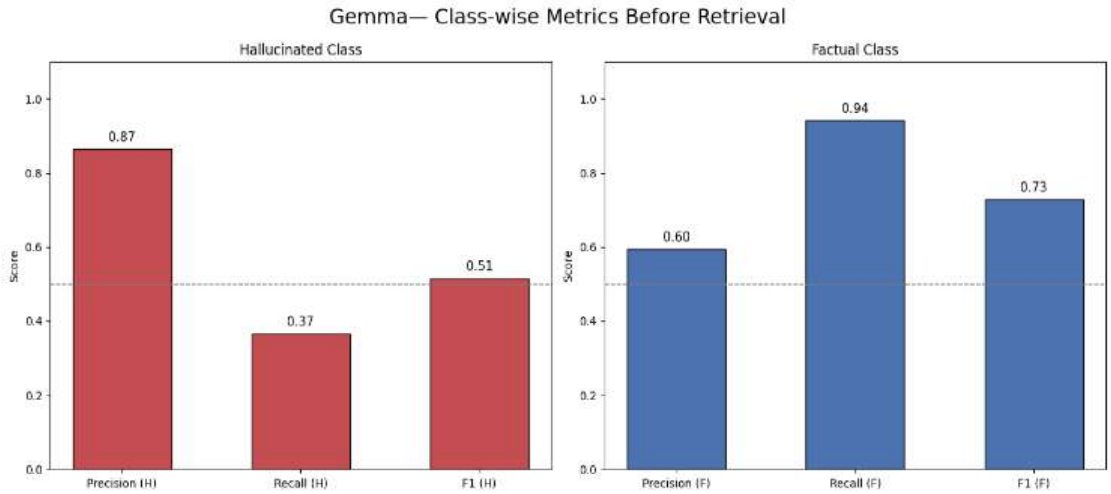


Figure 13: Class-wise metrics of Gemma 7B before retrieval

Gemma 7B achieved a precision of 0.87 recall of 0.37 and F1 score of 0.51 for the Hallucinated class. For the Factual class the precision was 0.60 recall was 0.94 and F1 score was 0.73 as shown in Figure 10. The lower Recall for the Hallucinated class indicates that a considerable number of hallucinated responses were not identified before retrieval.

B. Model-Detected Predictions after Retrieval

This subsection provides the detection output after the post-hoc retrieval pipeline, which decomposes misclassified answers into atomic claims, validates them against Wikipedia evidence, and re-labels them using FActScore.

GPT-3.5 Turbo:

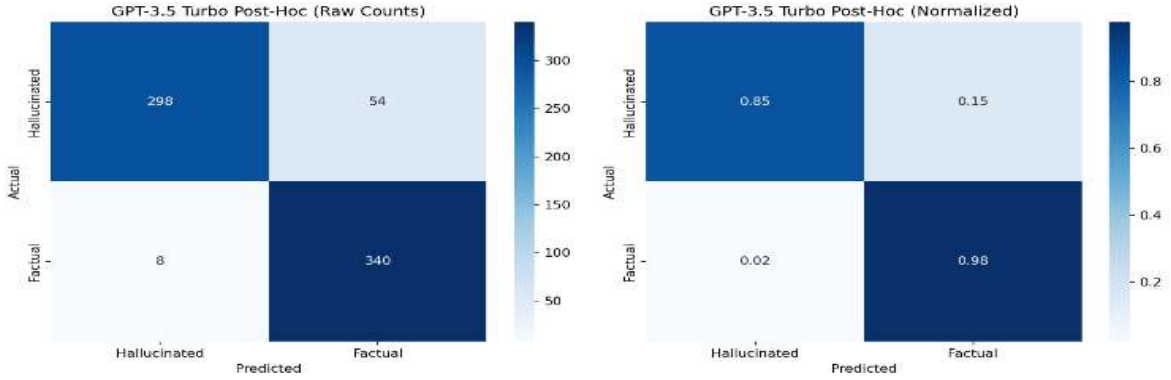


Figure 14: Confusion matrix of GPT-3.5 Turbo after evidence retrieval

After retrieval, GPT-3.5 Turbo correctly identified 298 of the 352 hallucinated responses (85%) and 340 of the 348 factual responses (98%), as shown in Figure 14. The detection of hallucinated responses increased from the pre-retrieval result to 85% after retrieval.

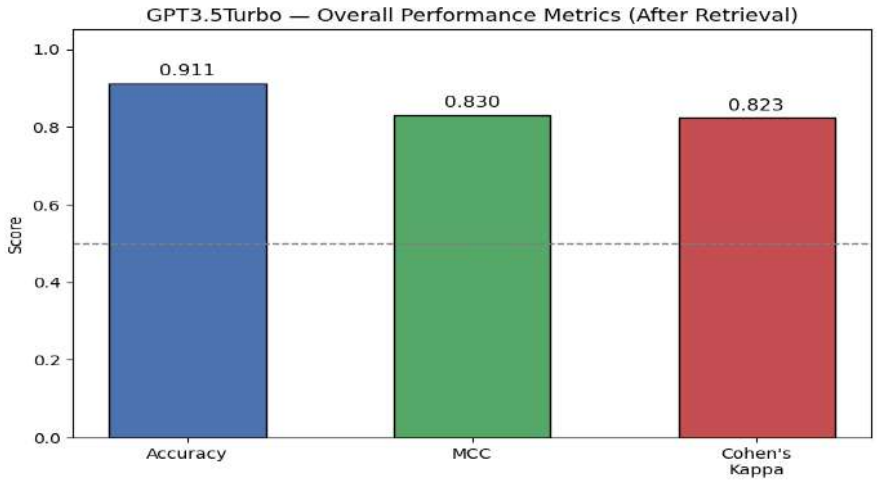


Figure 15: Overall performance metrics of GPT-3.5 Turbo after evidence retrieval

GPT-3.5 Turbo achieved an Accuracy of 0.911, MCC of 0.830, and Cohen's Kappa of 0.823 after retrieval, as shown in Figure 15, a substantial gain over the pre-retrieval results, particularly in agreement with ground truth.

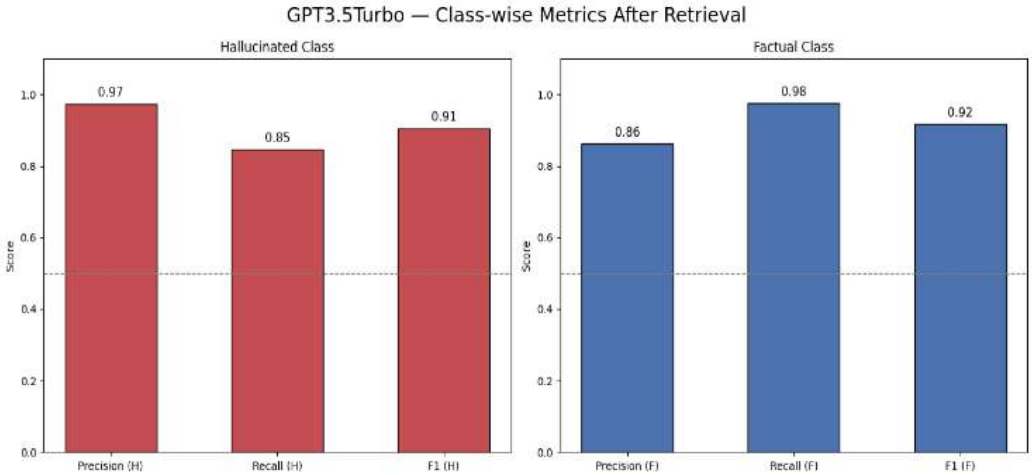


Figure 16: Class-wise metrics of GPT-3.5 Turbo after evidence retrieval

Recall for the Hallucinated class rose from 0.52 to 0.85 and F1 from 0.67 to 0.91, with precision remaining high at 0.97, as shown in Figure 16, showing that retrieval substantially improved detection while preserving precision.

GPT-4:

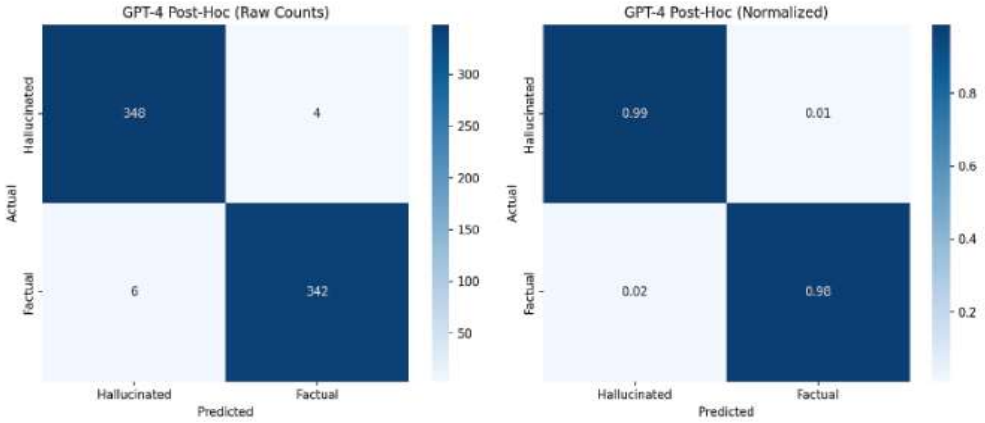


Figure 17: Confusion matrix of GPT-4 after evidence retrieval

GPT-4 achieved near-perfect performance after retrieval, correctly identifying 348/352 hallucinated (99%) and 342/348 factual (98%) responses, as shown in Figure 17, further improving its already strong detection capability.

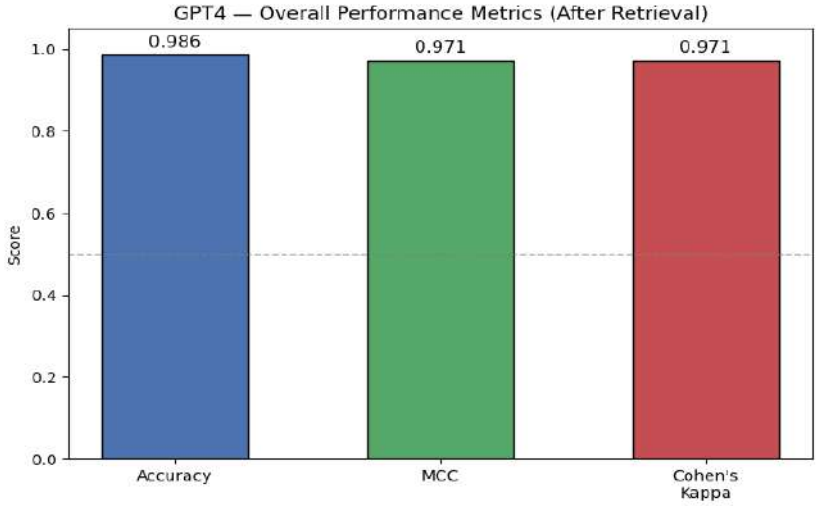


Figure 18: Overall performance metrics of GPT-4 after evidence retrieval

After retrieval, GPT-4's Accuracy reached 0.986, MCC rose from 0.844 to 0.971, and Cohen's Kappa from 0.837 to 0.971, as shown in Figure 18, reflecting very strong agreement between predicted and actual classifications.

Figure 19: Class-wise metrics of GPT-4 after evidence retrieval

After retrieval, GPT-4 achieved Precision, Recall, and F1-score values between 0.98 and 0.99 for both classes, as shown in Figure 19. The values for the two classes were very close, with only small differences across the three evaluation measures.

GPT-4 Turbo:

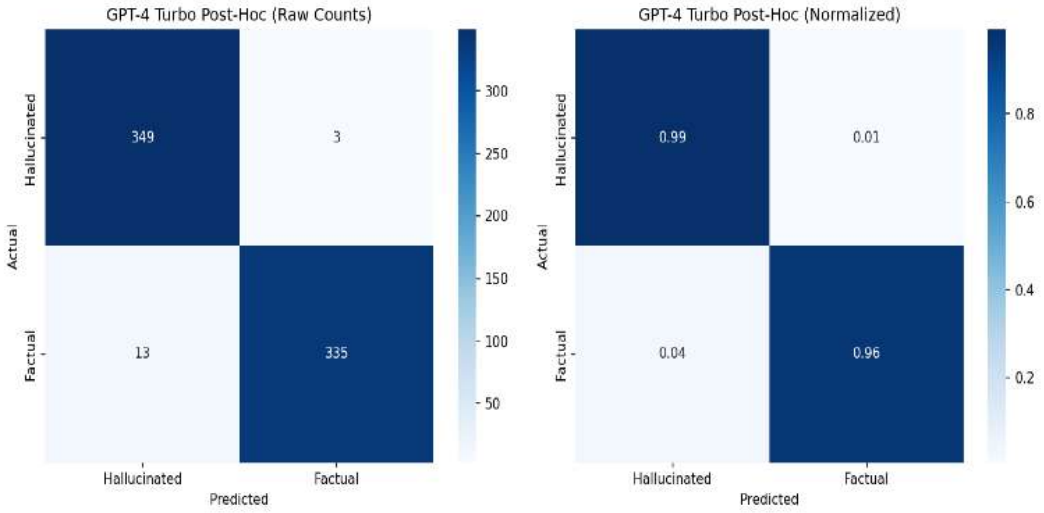


Figure 20: Confusion matrix of GPT-4 Turbo after evidence retrieval

GPT-4 Turbo, already the strongest pre-retrieval model, improved further after retrieval, correctly identifying 349/352 hallucinated (99%) and 335/348 factual (96%) responses, as shown in Figure 20.

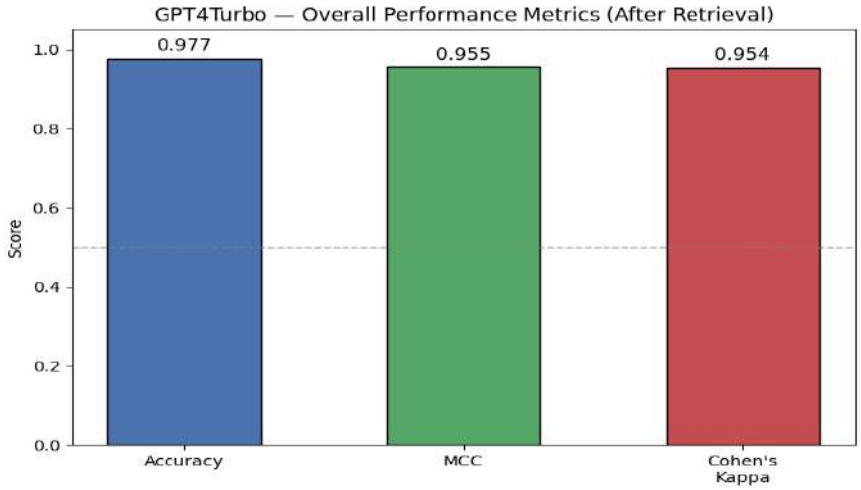


Figure 21: Overall performance metrics of GPT-4 Turbo after evidence retrieval

GPT-4 Turbo's Accuracy reached 0.977 after retrieval, with MCC rising from 0.886 to 0.955 and Cohen's Kappa from 0.886 to 0.954, as shown in Figure 21, confirming retrieval's benefit even for an already strong model.

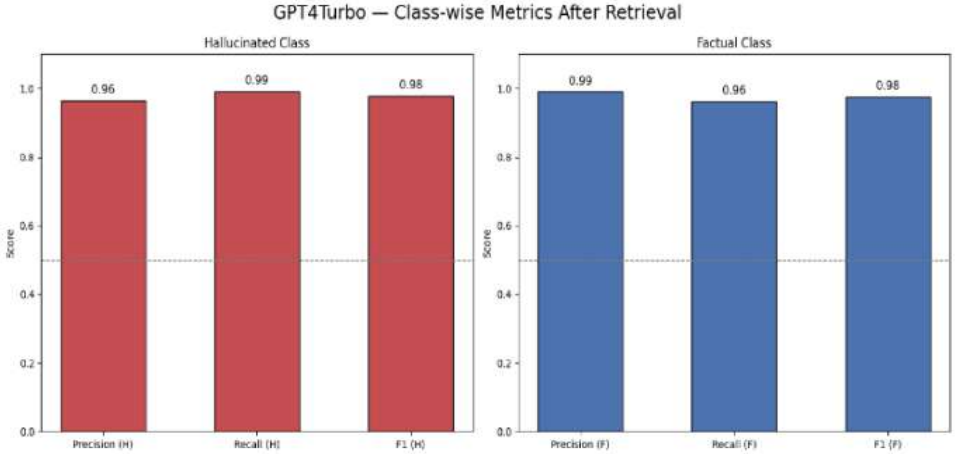


Figure 22: Class-wise metrics of GPT-4 Turbo after evidence retrieval

After retrieval GPT-4 Turbo got a Recall of 0.99 and an F1-score of 0.98 for the Hallucinated class. For the Factual class GPT-4 Turbo got a Precision of 0.99 a Recall of 0.96 and an F1-score of 0.98 as shown in Figure 22. The F1-score stayed at 0.98 for both the class and the Factual class while the Recall was a little lower, for the Factual class.

Gemma 7B:

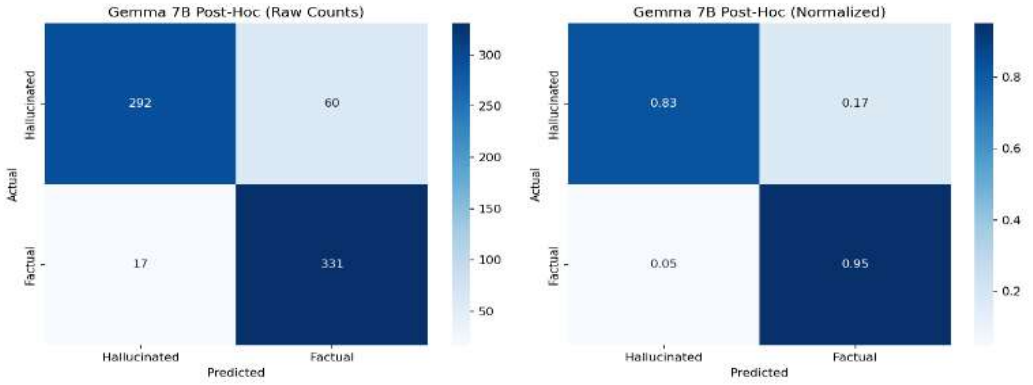


Figure 23: Confusion matrix of Gemma 7B after evidence retrieval

Gemma 7B showed the largest change among the four models after retrieval. The number of hallucinated responses correctly identified increased from 129 (37%) before retrieval to 292 (83%) after retrieval. At the same time, the number of factual responses incorrectly classified as hallucinated decreased from 20 to 17, as shown in Figure 23. Thus, retrieval substantially improved the identification of hallucinated responses for Gemma 7B.

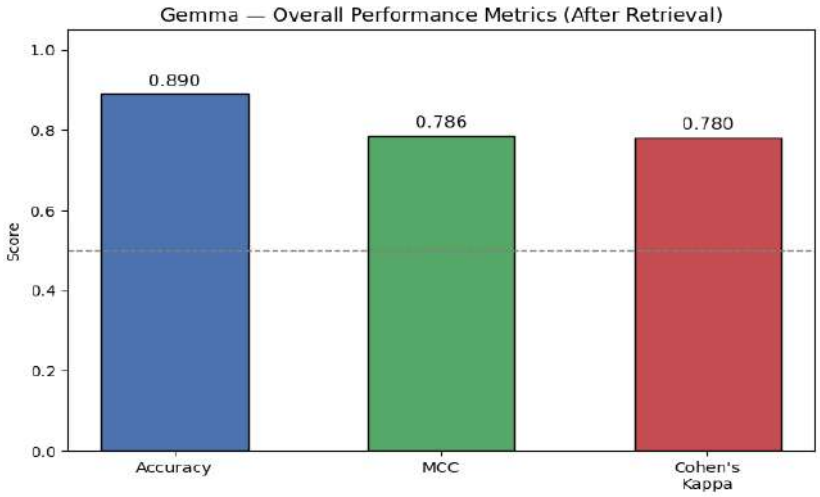


Figure 24: Overall performance metrics of Gemma 7B after evidence retrieval

After retrieval, Gemma 7B's Accuracy rose to 0.890 from 0.653, MCC from 0.377 to 0.786, and Cohen's Kappa from 0.308 to 0.780, as shown in Figure 24, marking a considerable gain in classification reliability.

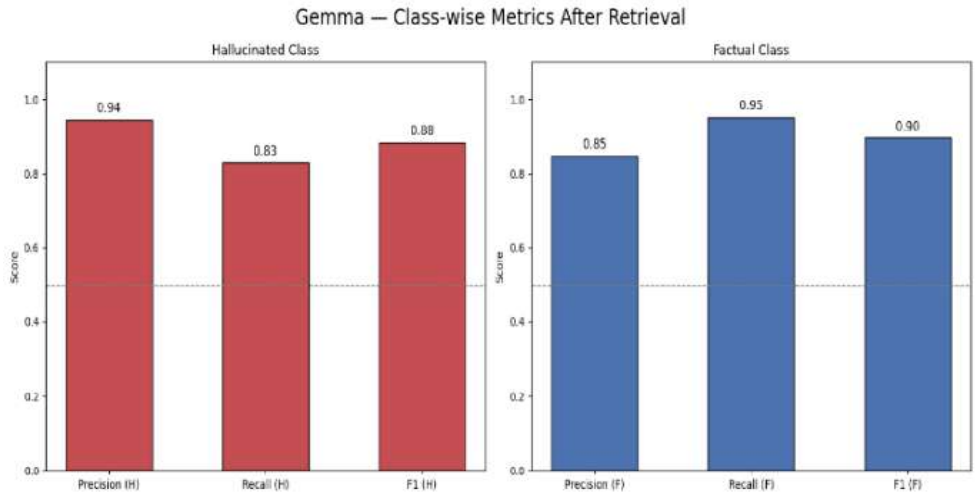


Figure 25: Class-wise metrics of Gemma 7B after evidence retrieval

Recall for the Hallucinated class rose from 0.37 to 0.83 and F1 from 0.51 to 0.88, with precision remaining high at 0.94, as shown in Figure 25, showing that retrieval substantially improved Gemma 7B's hallucination detection while preserving precision.

C. Inter-Model Agreement (Fleiss' Kappa)

Fleiss' Kappa was used along with Cohen's Kappa to measure agreement across the four LLMs. While Cohen's Kappa compares a model's predictions with the ground truth, Fleiss' Kappa measures agreement among all four models. Before retrieval, the Fleiss' Kappa value was 0.5438, indicating moderate agreement among GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, and Gemma 7B when using their internal knowledge. After retrieval, the value increased to 0.8289, showing that the classifications became more consistent when the models used the same external evidence for verification.

D. FActScore Threshold Analysis

The FActScore threshold determines the decision boundary for classifying responses as Factual or Hallucinated. Since detecting hallucinated information is the framework's primary goal, the threshold is chosen experimentally by analysing Hallucinated-class Precision and Recall across candidate FActScore values, selecting the value providing the best balance between the two for final classification.

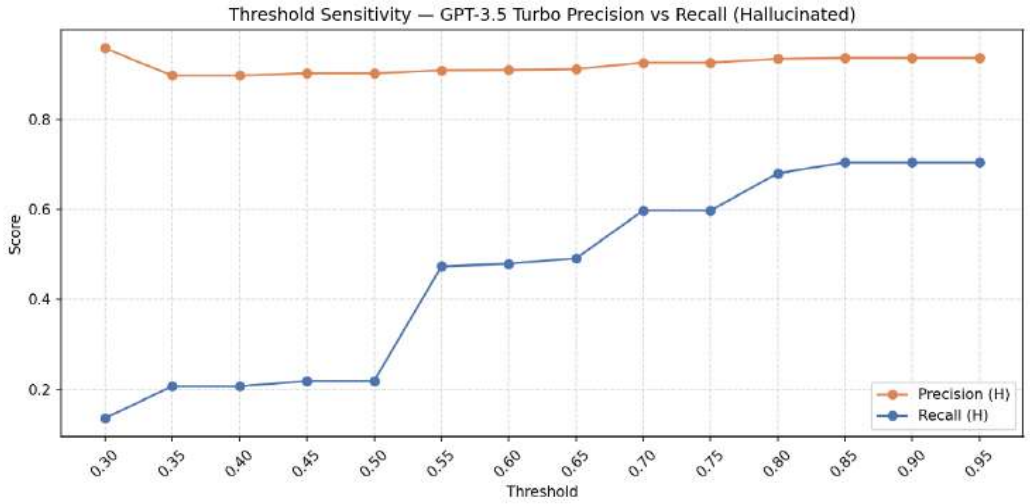


Figure 26: Threshold sensitivity analysis of GPT-3.5 Turbo based on Hallucinated-class Precision and Recall

As shown in Figure 26, for GPT-3.5 Turbo, the Precision is relatively stable under the tested thresholds, but the Recall is improved with the higher threshold. We selected a threshold of 0.85 as the best threshold for this model, which provides approximately 0.93 Precision and 0.70 Recall.

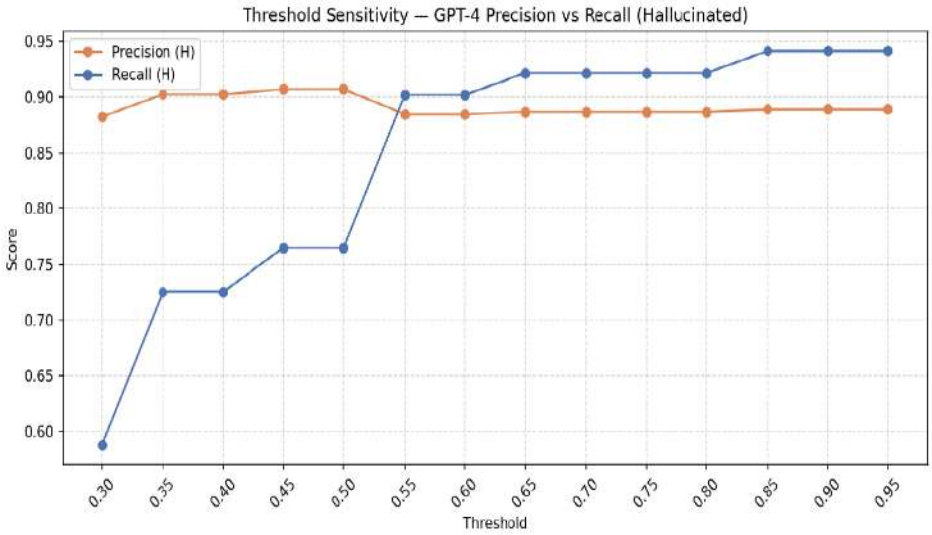


Figure 27: Threshold sensitivity analysis of GPT-4 based on Hallucinated-class Precision and Recall

For GPT-4 Precision stays relatively constant over the tested threshold values while Recall increases with the threshold as shown in Figure 27. A threshold of 0.55 the model achieves 0.89 Precision and 0.90 Recall For these numbers, the threshold was selected as 0.55 for GPT-4

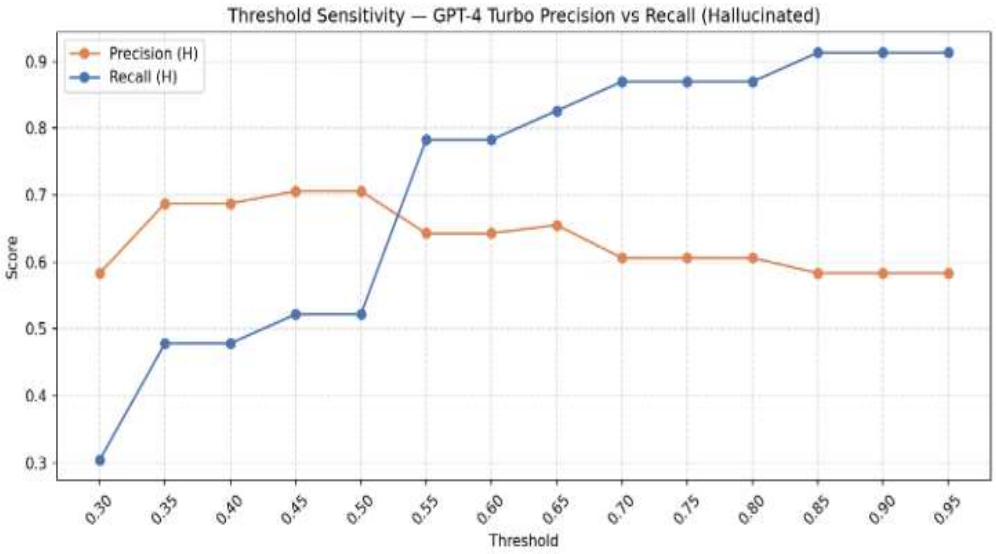
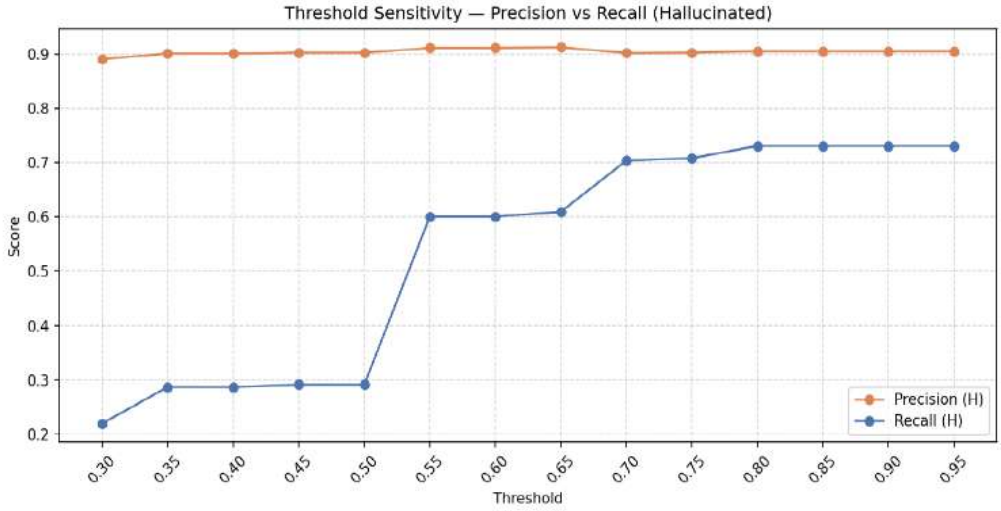


Figure 28: Threshold sensitivity analysis of GPT-4 Turbo based on Hallucinated-class Precision and Recall

For GPT-4 Turbo the Precision goes up. Reaches a high of 0.71 when the threshold is, between 0.45 and 0.50. The Recall was 0.52. The recall is around 0.78 at a threshold of 0.55. Precision is about 0.65. As shown in Figure 28 The Recall, at thresholds is about 0.91 but the Precision is at around 0.58. So, we set a threshold of 0.55 for GPT-4 Turbo



.

Figure 29: Threshold sensitivity analysis of Gemma 7B based on Hallucinated-class Precision and Recall

For the Hallucinated class, Precision remains around 0.89 to 0.91 across the tested thresholds, while Recall increases from approximately 0.22 at 0.30 to 0.60 at 0.55 and 0.73 at 0.80. Recall remains nearly unchanged at higher thresholds, as shown in Figure 29. A threshold of 0.80 was selected for Gemma 7B, with approximately 0.90 Precision and 0.73 Recall.

V. CONCLUSION AND FUTURE SCOPE

This study evaluated four LLMs for detecting factual hallucinations using 700 question and answer pairs from 20 academic domains. The results showed clear differences in their initial factuality detection performance. GPT-4 and GPT-4 Turbo performed very well on the current dataset even before retrieval, suggesting that the factual questions used in the evaluation were generally within the knowledge and capabilities of these advanced models. In contrast, Gemma 7B showed lower initial performance, which may be related to limited internal knowledge for some of the factual questions. After external evidence was retrieved and individual claims were verified, the detection results improved across all four models. The largest improvement was observed for Gemma 7B, where the identification of hallucinated responses increased from 37% to 83%. These findings show that external evidence can improve factual hallucination detection, particularly when the model's internal knowledge is not sufficient to make a reliable judgment.

Future work can test the proposed approach with a wider variety of factual questions instead of using only academic question datasets. Other factuality and hallucination datasets, together with real-world questions that contain more difficult or less obvious factual errors, can help assess how well advanced LLMs perform under more challenging

conditions. The dataset can also be expanded to include more domains and different types of factual information. Further research can explore different retrieval methods, domain-specific evidence sources, and adaptive FActScore thresholds. The current framework can also be extended to automatically correct unsupported factual claims using the retrieved evidence.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ACKNOWLEDGMENT

Authors express their sincere thanks to the management, faculty members and staff of Department of Computer Science and Engineering, UCEK, JNTUK, Kakinada for their valuable support, guidance, encouragement and cooperation throughout this research work.
Authors thank the reviewers for their valuable comments and constructive suggestions to improve the quality and clarity of the manuscript.

REFERENCES

[1] R. H. Ajmal, M. U. Sarwar, M. K. Hanif, and M. I. Khan, "Evaluating the Effectiveness of Advanced Language Models in Detecting and Mitigating Hallucinations Using Structured Question-Answering, Novel Metrics, and Post-Hoc Retrieval," *IEEE Access*, vol. 13, pp. 173805–173815, 2025. Available from: https://ieeexplore.ieee.org/document/11177188

[2] Saxena, D. Aggarwal, N. Badathala, and P. Bhattacharyya, "A framework to synthetically generate fine-grained hallucinated data," *Language Resources and Evaluation*, vol. 59, pp. 423–452, 2025. Available from: https://link.springer.com/article/10.1007/s10579-025-09864-x

[3] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models," in *Proc. 2023 Conf. Empirical Methods in Natural Language Processing*, 2023, pp. 6449–6464. Available from: https://aclanthology.org/2023.emnlp-main.397/

[4] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," arXiv:2303.08896, 2023. Available from: https://arxiv.org/abs/2303.08896

[5] Y. Zhang, "Hallucination Detection in Large Language Models by Classifying Information Sources," in *Proc. CONF-SPML 2026 Symposium: 2nd Neural Computing and Applications Workshop*, 2025. Available from: https://doi.org/10.54254/2755-2721/2026.TJ30007

[6] S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12076–12100, 2023. Available from: https://aclanthology.org/2023.emnlp-main.741/

[7] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers)*, 2022, pp. 3214–3252. Available from: https://aclanthology.org/2022.acl-long.229/

[8] Z. Cao, Y. Yang, X. Li, and H. Zhao, "AutoHall: Automated Factuality Hallucination Dataset Generation for Large Language Models," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 34, 2026. Available from: https://ieeexplore.ieee.org/document/11267179/

[9] Abdeen, S. M. T. Siddiqui, M. T. Ahmed, A. Singhal, and L. Khan, "Hallucination Detection in Large Language Models Using Diversion Decoding," Univ. of Texas at Dallas / NIST, 2024. Available from: https://doi.org/10.1007/978-3-031-96590-6_7

[10] Barathidass and K. Namburu, "Rapid End-to-End Test Generation and Hallucination Mitigation Using Generative Artificial Intelligence," *IEEE Access*, vol. 14, 2026. Available from: https://doi.org/10.1109/ACCESS.2026.3657407

[11] B. Zhu, X. Zhang, M. Gu, and Y. Deng, "Knowledge Enhanced Fact Checking and Verification," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 29, 2021. Available from: https://doi.org/10.1109/TASLP.2021.3120636

[12] Z. Lin, S. Guan, W. Zhang, H. Zhang, Y. Li, and H. Zhang, "Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models," *Artificial Intelligence Review*, vol. 57, no. 243, 2024. Available from: https://link.springer.com/article/10.1007/s10462-024-10896-y

[13] S. S. Rahman et al., "Hallucination to truth: a review of fact-checking and factuality evaluation in large language models," *Artificial Intelligence Review*, vol. 59, no. 70, 2026. Available from: https://link.springer.com/article/10.1007/s10462-025-11454-w

[14] N. Sharma and S. Singh, "A Hybrid Extractive and Encoder-Decoder-Based Approach for Mitigating Hallucination in Automatic Text Summarization," *J. Transformative Technologies and Sustainable Development*, vol. 9, no. 15, 2025. Available from: https://link.springer.com/article/10.1007/s41314-025-00083-4

[15] S. Liu, Y. Gao, S. Li, P. Wang, and T. Wang, "A hallucination detection and mitigation framework for faithful text summarization using LLMs," *Scientific Reports*, vol. 16, no. 1374, 2026. Available from: https://www.nature.com/articles/s41598-025-31075-1

[16] M. Omar et al., "Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support," *Communications Medicine*, vol. 5, no. 330, 2025. Available from: https://doi.org/10.1038/s43856-025-01021-3

[17] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023. Available from: https://doi.org/10.1145/3571730

[18] H.-Y. Lin, Y.-C. Chang, and J.-T. Chien, "Variational Dehallucination via Retrieval-Augmented Prompt Learning," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 34, pp. 259–271, 2026. Available from: https://doi.org/10.1109/TASLPRO.2025.3646477

[19] L. Kong, X. Zhong, J. Chen, H. Fu, and Y. Wang, "Multi-perspective consistency checking for large language model hallucination detection: a black-box zero-resource approach," *Frontiers of Information Technology & Electronic Engineering*, vol. 26, no. 11, pp. 2298–2309, 2025. Available from: https://www.fitee.zjujournals.com/en/article/doi/10.1631/FITEE.2500180/

[20] H. Huang, X. Bu, H. Zhou, Y. Qu, J. Liu, M. Yang, B. Xu, and T. Zhao, "An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4," in *Findings of ACL 2025*, Vienna, Austria, 2025, pp. 5880–5895. Available from: https://aclanthology.org/2025.findings-acl.306/