Correcting Factuality Hallucination in Complaint Large Language Model via Entity-Augmented

Jiaju Kang[1], Weichao Pan[1], Tian Zhang[2], Ziming Wang[1],
Shuqin Yang[1], Zhiqin Wang[1], Jian Wang[1], Xiaofei Niu[1]*
[1]School of Computer Science and Technology, Shandong Jianzhu University,
1000 Fengming Road , Jinan, 250101, Shandong, China.
[2]Ecole supérieure d'ingénieurs en génie électrique,
Technopôle du Madrillet, Av. Galilée, 76800 Saint-Étienne-du-Rouvray, France.
xiaofein@sdjzu.edu.cn; 202011110193@stu.sdjzu.edu.cn

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10650208

***Abstract*—Complaint Large Language Model (Complaint-LLM) is designed as a "customer service" tool to address the scenario of handling a massive volume of public complaints, effectively leveraging the "common sense" possessed by Large Language Models (LLMs) to solve issues. Unfortunately, pre-trained LLMs often exhibit significant Factual Hallucination and Causal Errors in knowledge domains with sparse experience distribution, greatly affecting the accuracy of user interactions with LLMs. We propose an architecture that utilizes external data to support pre-trained models, aiming to avoid the expensive cost of retraining LLMs. The core concept involves leveraging prompts to inject strongly correlated additional information into LLMs and adjusting the initialized alternative outputs along the inference pathway of the LLM. To achieve this, we construct a rich knowledge graph as a knowledge base for algorithm retrieval and learning. Each input text is decomposed into subgraphs corresponding to nodes on the knowledge graph, and a graph neural network classifier is trained to obtain classification results and additional knowledge. Numerous experiments demonstrate that the Complaint-LLMs shows a significant improvement in the question-answering evaluation of various subclass scenarios in the complaint domain. Moreover, the graph neural network trained with complaint text data exhibits good transferability in classification tests for open scenarios.**

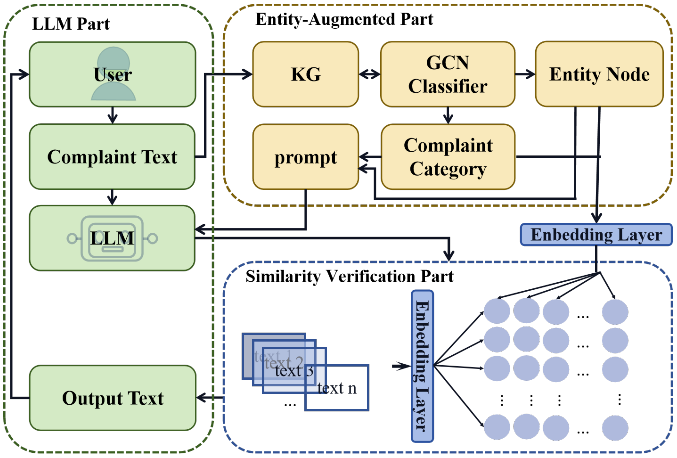


Fig. 1. We propose a processing flow that uses external knowledge graph(KG) and pre-trained graph neural networks(GCN Classifier) to provide additional knowledge for LLM, with a set of elegantly designed prompt to support model computation and constrain model output. The alternative solutions generated by the model will undergo a similarity check with retrieved entity knowledge using a pre-trained Bert model, yielding a top-k sequence. The model will output the optimal solution after evaluating the confidence level of the assessment results and consistency metrics.

I. Introduction

The advent of trendsetting language models, including T5 [1], GPT-3 [2], LLaMA [3], Gemini [4], and others, has marked a pinnacle in the realms of Natural Language Processing (NLP) and Human-machine Dialogue. Compared with previous models, LLMs have achieved surprising results in open-domain question answering tasks, thanks to their rich training data and the ability to learn from context based on transformers. In dialogue scenarios, their coherent and appropriate responses make them suitable for direct communication with users as customer service agents, some works [5], [6] have demonstrated the potential of LLM as chatbot agents. However, in domains with specific knowledge, such as complaint scenarios, when the information provided by the user is more fine-grained, LLM are limited by their systematic generalization ability, and often exhibit significant factuality hallucination and causal errors. This is reflected in the dialogue box as irrelevant or erroneous text, which is called the hallucination of LLM. Entity enhancement is an effective method to correct low-quality model outputs caused by the lack of relevant knowledge. One useful method is to use prompt embedding statements to give LLM extra experience to cope with complex interaction scenarios.

Essentially, method of this kind are entity-level retrieval-enhanced based on external knowledge bases, providing LLMs with extra information and rich reasoning logic that are not in the training data [7]. A similar work, [8] integrated extra knowledge into the decoder of a CodeT5-base model, thereby improving the code generation ability. However, Complaint-LLM faces a wide and diverse range of discussion topics, and we hope that the entire discussion process is user-friendly, which means that the model needs to consider the whole context more and precisely control the discussion content within this scope. Therefore, before embedding the users input,

*Corresponding author

Authorized licensed use limited to: Vignana Bharathi Institute Of Technology. Downloaded on August 05,2025 at 06:02:09 UTC from IEEE Xplore. Restrictions apply.

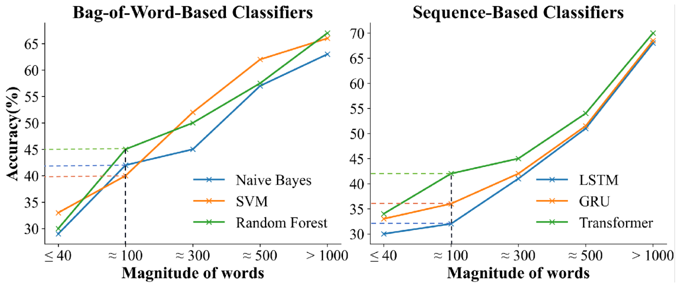


Fig. 2. We observed the impact of text length on classifier performance. Specifically, using GPT4, we randomly augmented a short text dataset containing 100 data samples to create five datasets with different volumes but the same semantics. In the traditional classification approaches considering sentences as sequences or bag-of-words, we observed a clear increase in classification accuracy as the individual text length grew. Additionally, we found that the knowledge abstraction of traditional methods did not perform well on short text datasets in such attempts. Complaint short texts typically have a word count of around 100 words.

LLM will additionally receive a set of dedicated prompts to regulate the thinking chain of LLM, such as the category of users needs, the relevant information of the government departments that can be contacted to solve the users needs, the actual handling process and feedback, format constraints for model output etc., so that the model can fully think about the problem being discussed.

Retrieving the category and relevant information of a complaint text from a large-scale knowledge base is a classic short text classification problem. Previous work has found that, because complaint short texts contain less details and have a colloquial narrative style, they produce unacceptable errors in multi-label classification tasks [9]. For example, "<USER> reported that there was a problem of illegal private occupation of the elevator in the Victory Garden Community. The staff arrived at the scene and found that there were unregistered users using it illegally and shut it down. Now it is said that the elevator is still running after <DATE>, and requests that all relevant personnel involved in the process be investigated, otherwise it will be exposed to the media." The accurate positive sample labels should include Illegal use of Special Equipment, Poor Handling by Relevant Personnel, and Media Public Relations Risk as three important categories, involving the Bureau of Industry and Commerce, the community property, the Personnel Bureau and the Public Relations Department.The typical length of complaint texts is around 100 words, involving numerous associated entities and at least two classification labels that fully conform to the semantic meaning of the text. Therefore, we contemplated whether traditional classification models could yield acceptable results.

In Figure 2, we aimed to investigate the classification performance of traditional models on complaint texts with varying data volumes but the same topic. Both extensively fine-tuned bag-of-words and sequence-based models on the complaint text dataset did not achieve the expected experimental outcomes. The approach of entity augmentation using traditional methods encountered a stalemate.

Knowledge graph as a knowledge repository, coupled with a graph neural network serving as a classifier, has captured our attention as a retrieval solution [10]. Intuitively, on the one hand, a knowledge graph stores textual knowledge in a structured format, exhibiting robust data update capabilities, which are valuable for alleviating the high training costs associated with LLM. This implies that the model need not undergo retraining and fine-tuning from larger datasets to gain new insights. On the other hand, knowledge retrieval based on verifiable information allows quantifying the uncertainty of generated results using confidence or consistency metrics under controlled conditions. Experimental results demonstrate that retrieval enhancement based on Knowledge Graph Convolutional Networks (KGCN) exhibits exceptional performance in this context. Indeed, with the integration of an efficient and accurate adaptive graph neural network classifier and retrieval strategy, the four-layer ontology-based knowledge graph developed in our work is adept at handling the intricate communication tasks in the context of complaint scenarios.

The main contributions of our work are summarized as follows:

- Within the scope of our investigation, we pioneered the use of a retrieval enhancement method based on KGCN to extend the commendable performance of LLM into the realm of automated complaint processing.
- We release a new large-scale Chinese complaint dataset, annotated using government standards. Based on this, we generate an open-source complaint knowledge graph (Complaint-KG).
- We reevaluated the impact of the data scale for knowledge injection on the mitigation of LLM hallucination.

II. Raleted Work

Hallucination: The hallucination in LLM poses a significant factor restricting users from obtaining accurate and efficient knowledge from the feedback of the LLM [11]. This phenomenon may arise due to the use of erroneous, low-quality data during LLM training [12], and the repeated emphasis on certain data inputs may result in memory biases in such situations [13], [14]. Alternatively, during logical reasoning, an incorrect assertion may persist from a minor error, leading to inaccurate final outputs [15].

Retrieval Enhancement: Two essential retrieval strategies significantly enhance the generative performance of the model from different perspectives [16]. The first is pre-input retrieval, where an external module extracts relevant knowledge from specialized knowledge bases before the model reads the initial text, injecting this information into the LLM through prompts [17]–[20]. The second is post-output retrieval, typically occurring after decoding [21]–[23]. The model generates a set of alternative solutions, and an external module compares the generated text with the extracted external knowledge, selecting the most effective result to reduce the risk of illusions.

III. Method

In this section, we will introduce an external module designed for correcting model illusions. This module is a

Authorized licensed use limited to: Vignana Bharathi Institute Of Technology. Downloaded on August 05,2025 at 06:02:09 UTC from IEEE Xplore. Restrictions apply.

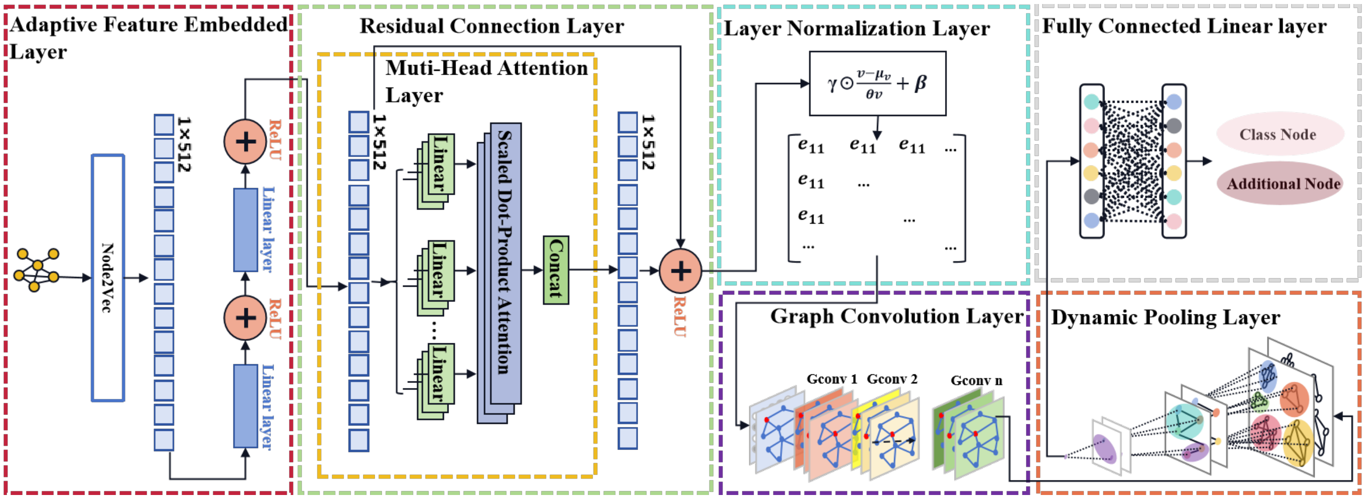
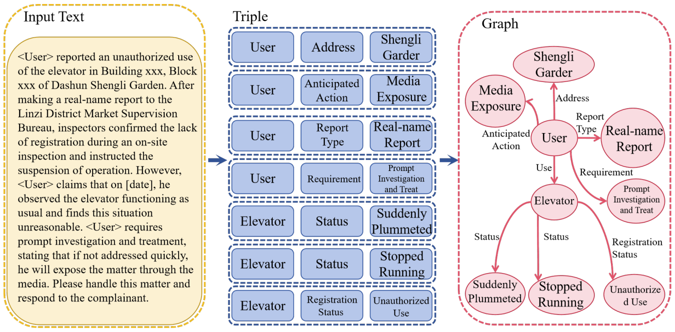

Adaptive Feature Embedded Layer
Node2Vec
1×512
Linear layer
ReLU
Linear layer
ReLU
Residual Connection Layer
Muti-Head Attention Layer
1×512
Linear
Linear
Linear
Scaled Dot-Product Attention
Concat
1×512
ReLU
Layer Normalization Layer
$\gamma \odot \frac{v-\mu_v}{\theta v} + \beta$
e_{11} e_{11} e_{11} ...
e_{11} ...
e_{11} ...
... ...
Fully Connected Linear layer
Class Node
Additional Node
Graph Convolution Layer
Gconv 1
Gconv 2
Gconv n
Dynamic Pooling Layer

Fig. 3. The overall architecture of our proposed KGCN, , designed to provide accurate class nodes and additional information nodes for a given subgraph.



<User> reported an unauthorized use of the elevator in Building xxx, Block xxx of Dashun Shengli Garden. After making a real-name report to the Linzi District Market Supervision Bureau, inspectors confirmed the lack of registration during an on-site inspection and instructed the suspension of operation. However, <User> claims that on [date], he observed the elevator functioning as usual and finds this situation unreasonable. <User> requires prompt investigation and treatment, stating that if not addressed quickly, he will expose the matter through the media. Please handle this matter and respond to the complainant.

User	Address	Shengli Garder
User	Anticipated Action	Media Exposure
User	Report Type	Real-name Report
User	Requirement	Prompt Investigation and Treat
Elevator	Status	Suddenly Plummeted
Elevator	Status	Stopped Running
Elevator	Registration Status	Unauthorized Use

Fig. 4. For each input complaint text, the model extracts and transforms it into a connected subgraph that represents the central semantics.

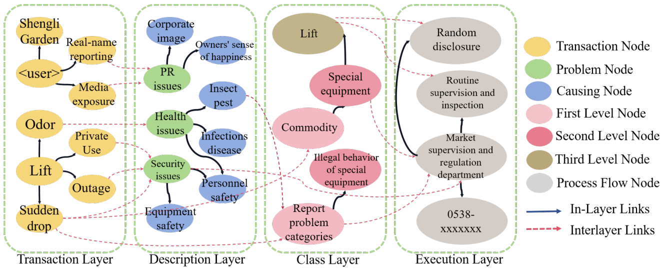


Fig. 5. Complaint-KG. Utilizing a four-layer knowledge graph structure, we illustrate the linkages of the sample text on the knowledge graph.

retrieval validation module based on knowledge graphs and graph neural networks. The entire workflow, as depicted in Figure 1, unfolds as follows:

- Upon receiving a user complaint ticket, the Entity Extraction module processes it into a graph format, as illustrated in Figure 4.
- The KGCN module retrieves the category of this graph and provides relevant entities, referred to as "additional knowledge."
- The additional knowledge, along with a prompt and instructions representing constraints, is input into the model.
- The model generates multiple candidate solutions, and we use a pre-trained BERT model based on cosine similarity to systematically assess the correlation between the generated text and additional knowledge.
- The best generating result is then output.

The automated process eliminates the need for users to perform any additional actions to adjust the model. This method seamlessly integrates knowledge graphs, graph neural networks, and pre-trained language models, offering an effective approach.

A. Class Entity Retrieval

The purpose of class entity retrieval is to assign multiple class labels to unfamiliar text based on existing nodes in the knowledge graph. We designed a KGCN similar to BERT, in which a Graph Convolution Layer and a Dynamic Pooling Layer are bridged at the end of the network. A pre-trained Adaptive Graph Neural Network is employed to learn relationships between nodes across the entire knowledge graph, providing precise classification labels for input subgraphs.

As shown in Figure 5, the Complaint-KG consists of four well-designed layers. The Transaction Layer encompasses all entities and relationships that may occur in a complaint, providing a refined overview of the entire complaint scenario. The Description Layer further analyzes the potential consequences of events within the Transaction Layer, serving as additional logic to guide the linkage between complaint text and categories. The Class Layer utilizes the official classification system of government departments and is designed as an immutable layer with a fixed number of nodes. The Execution Layer contains all relevant information about the government department corresponding to the complaint, sourced from publicly available data on government websites. We observe a strong correlation between the Class Layer and the Execution Layer, and their adjacency yields the most effective explanatory results. Rich linkages exist between layers, extending beyond the concept of ’adjacent layers.

Authorized licensed use limited to: Vignana Bharathi Institute Of Technology. Downloaded on August 05,2025 at 06:02:09 UTC from IEEE Xplore. Restrictions apply.

B. Additional Entity Retrieval

Using only class labels for text input annotations has limitations in enhancing model understanding. Therefore, based on the subgraph generated from the input text, we perform backtracking on the knowledge graph. We employ a semi-random stepwise expansion operation to increase the number of entity nodes in the entire subgraph. Specifically, at critical nodes where the feature distance reaches a certain threshold, we randomly choose an expansion direction. This signifies a concept aimed at identifying adjacent nodes with important semantic relationships to the node and adding them to the algorithm's receptive field.

During the actual computation, we aim for the algorithm to focus on entity nodes in the Transaction Layer, Description Layer, and especially the Execution Layer when seeking additional nodes. When the algorithm determines the relevance of the complaint text to a government department, all leaf nodes of that provincial government department will be outputted as much as possible.

We introduce a module based on Information Gain (IG) and Node Importance (Imp) to determine whether newly added nodes contribute to the knowledge representation of the entire subgraph. When this module cannot decide whether to include the remaining nodes in the subgraph, the random selection process is repeated. We will delve into the use of these two key factors in the process.

IG: We can calculate the information gain by treating the clustering coefficient of a graph as entropy. The clustering coefficient describes the degree of clustering of nodes in the graph and is defined as follows:

$$C = 3 * Num_{\text{triangle}} / Num_{\text{triples}} \tag{1}$$

Where Num_{triangle} is the number of all fully connected triangles (formed by three nodes) in the graph, and Num_{triples} is the number of all connected triplets (not necessarily fully connected) in the graph.

Assuming that the clustering coefficient of the current subgraph is C_{before}, and the clustering coefficient of the subgraph after adding the new node N is C_{after} , the information gain $IG(N)$ can be defined as the difference between C_{before} and C_{after} , i.e

$$IG(N) = C_{\text{before}} - C_{\text{after}} \tag{2}$$

Imp: We can define node importance as near centrality (Closeness Centrality). Proximity centrality is a measure describing the central position of a node in a network, defined as the mean of the inverse of the shortest distances from all nodes to that node.

For graph $G = (V, E)$, the near centrality $C(v)$ of node v is calculated as follows:

$$C(v) = \sum_{u \neq v \neq w} \frac{1}{d_{uv}} \tag{3}$$

Where $C(v)$ represents the near-centrality of node v, d_{uv} represents the shortest path length between node u and node v, and w represents the set of other nodes in the graph.

Thus, for node N , the importance is its near-centrality in the graph, i.e

$$Imp(N) = C(N) \tag{4}$$

Then, we can calculate the score of each node to be added through a weighted function. This score determines which node should be considered to be added to the subgraph:

$$Score(N) = w_1 * IG(N) + w_2 * Imp(N) \tag{5}$$

C. Prompts Engineering

We meticulously crafted a prompt set to more effectively integrate constraint conditions, complaint texts, and category labels obtained through model retrieval, along with additional knowledge. We considered various scenarios and complexities, ensuring that the cue words maintain accuracy and consistency while conveying information. This design encompasses not only the deep integration of domain-specific knowledge but also addresses the specific requirements of language models, providing LLM with richer input information. In our design, we placed particular emphasis on the precise expression of constraint conditions, ensuring that LLM can thoroughly consider specific limitations and requirements when understanding inputs. Simultaneously, the introduction of complaint texts aims to endow the model with enhanced emotional understanding and situational awareness, enabling it to interpret user intent more comprehensively. The embedding of category labels retrieved through model inference and additional knowledge provides LLM with a broader background, imbuing it with greater depth and breadth in language generation.

D. Candidate Verification

Despite providing extensive additional knowledge for LLM, ensuring the authenticity and relevance of black-box model outputs remains challenging. Therefore, we propose the adoption of a pre-trained cosine similarity module to assess the relationship between the node knowledge acquired in previous steps and candidate texts. In this approach, node data is cleverly embedded into a cue template, processed alongside candidate texts through a BERT encoder. Ultimately, the system will output results that are most relevant to and scored highest with the node knowledge. A particular scenario arises when the similarity between all generated texts and additional knowledge falls below a reasonable threshold, prompting the model to reject requests beyond its capabilities.

IV. Database

A. GCD Dataset

In order to promote research on Chinese short-text classification for complaints, we have formed the Government Complaints Data (GCD) dataset based on collected real complaint data. This dataset has the following characteristics:

Authorized licensed use limited to: Vignana Bharathi Institute Of Technology. Downloaded on August 05,2025 at 06:02:09 UTC from IEEE Xplore. Restrictions apply.

Diversity of Scenarios: The GCD dataset covers multiple major categories. Specifically, each text is assigned three class labels: Appeal Type, Appeal Classification, and Object Classification. Appeal types include reporting, complaints, inquiries, and requests for help. Appeal classifications consist of 3 primary classes, 47 secondary classes, and 133 tertiary classes. Object classifications include 2 primary classes, 50 secondary classes, and 271 tertiary classes.

TABLE I
DISTRIBUTION OF SCENARIOS IN THE GCD DATASET

Dataset	Data Volume	Description of Scenarios
GCD-FSC	1097	Food quality or safety issues
GCD-QC	929	Product quality issues
GCD-CC	454	Problems or disputes in the performance of the contract
GCD-ASC	534	Quality of service issues encountered after purchasing a product
GCD-UCC	179	Business practices that violate the principle of fair competition
Other Categories	1119	Complaint data in other scenarios

Multimodal data: In addition to the original text, for each topic in the dataset, we provide a dataset organized in a triplet format corresponding to the text in the GCD dataset, allowing users to evaluate the models extraction ability and accuracy. The generation of these triplets involves a small portion from chatGPTs understanding, but a larger amount of work comes from manual correction and relation extraction models based on joint decoding.

In previous research, we did not find high-quality complaint text data specifically targeted at government departments. Therefore, in model validation, we conducted extensive experiments on the collected GCD dataset. In addition to the 4,312 labeled texts contained in the GCD dataset, we have also made publicly available 4,162 unlabeled texts, along with a triple dataset generated from these texts.

The experiments mainly used four major components from the GCD dataset, namely GCD-FSC (GCD - Food Safety Complaints), GCD-QC (GCD - Quality Complaints), GCD-CC (GCD - Contract Complaints), and GCD-ASC (GCD - After-sales Service Complaints). These components focus on the four most common data scenarios in government complaint domains. They were manually classified by hotline operators in government complaint departments using speech-to-text conversion technology. We cleaned, corrected, filtered, and refined the original complex data to extract four relevant topics that reflect the textual situation in real complaint scenarios. In the data, we provide the original complaint text and three labels including demand type, demand classification, and object classification. The category space of data labels can be found in a separate file. The data comes from real texts received by government departments from January 2021 to September 2022.

B. Q & A Dataset

We utilized our processing framework to evaluate three datasets with varying levels of reasoning difficulty, providing a detailed report on the model's EM and F1 scores on these datasets. These metrics are designed to objectively reflect the model's performance in a question-answering scenario. Specifically, we covered three datasets: Simple Question [24], Mintaka [25], and HotpotQA [26].

V. EXPERIMENTS

A. LLM and KG Implementation

We employ a close-source model, ChatGPT (GPT-3.5) [27], and a pre-trained model, LLaMA(LLama-2-70B-Chat) [3], as the interface for LLM. The objective is to thoroughly evaluate whether integrating the pre-trained LLM into our framework exhibits significant performance improvements. Whether our focus is on Q&A datasets or the GCD dataset we collected, we consistently utilize the well-designed KG-Complaint, as depicted in Figure 5.

B. Baseline

We compared our proposed framework with the following three methods [20]:

- **Vanilla**: This employs a straightforward approach, prompting the model to generate answers directly for a given question.
- **Chain of Thought (CoT)** [28]: The objective is to generate more reliable answers by instructing the LLM to produce more comprehensive and detailed explanations for the generated answers.
- **Question Knowledge Retrieval (QKR)** [29]: This method prompts the LLM to generate answers using facts retrieved from a knowledge graph that are relevant to the given question. In this context, our objective is to demonstrate that, by incorporating the Candidate Verification Module, our approach is more effective compared to simply using knowledge graph enhancement methods alone.

C. Evaluation

We evaluated the accuracy and F1 score of the near-source model ChatGPT and the pre-trained LLaMA on the GCD dataset for the classification task. Across four open Q&A datasets, we assessed the EM and F1 scores of the models under tasks with varying levels of reasoning difficulty. In contrast to our processing pipeline, the three control processes considered were: Vanilla, CoT, and QKR.

It is worth noting that the performance improvement of LLM does not show significant contrast when injected knowledge is based on retrieval, as opposed to not being based on retrieval [30]. In other words, in experiments, the model can be strengthened even with noise data unrelated to the questions, yielding better results. The author proposes that this ambiguity may arise from the LLM treating injected knowledge as noise. The addition of extra knowledge length might even deteriorate the model's output. Our experiment aims to verify how the

Authorized licensed use limited to: Vignana Bharathi Institute Of Technology. Downloaded on August 05,2025 at 06:02:09 UTC from IEEE Xplore. Restrictions apply.

TABLE II
COMPARATIVE EXPERIMENTS ON Q & A DATASETS(EM AND F1)

		Simple Question		Mintaka		HotpotQA	
		ChatGPT	LLaMA	ChatGPT	LLaMA	ChatGPT	LLaMA
Vanilla		22.90/29.52	36.00/45.47	41.09/53.95	38.36/42.87	20.32/31.59	22.89/33.85
CoT		10.81/12.50	42.86/49.58	**54.86**/50.31	42.20/53.84	20.31/32.46	22.48/36.24
QKR		53.03/59.34	58.64/61.12	49.13/51.90	40.27/53.69	23.29/37.24	22.24/35.92
ours	@0.1	**59.14/62.34**	54.86/**64.99**	51.46/52.91	40.02/**54.44**	27.15/37.45	21.07/**44.6**
	@0.5	58.07/60.10	**61.02**/60.92	53.85/**54.60**	**47.12**/51.86	**27.30**/39.52	**23.30**/40.62
	@1	54.43/58.82	56.88/63.42	49.63/52.60	41.70/54.31	24.88/**40.66**	23.10/31.83

TABLE III
COMPARATIVE EXPERIMENTS ON MULTI-SCENE GCD DATASET(ACCURACY AND F1)

		GCD-FSC		GCD-QC		GCD-CC		GCD-ASC	
		ChatGPT	LLaMA	ChatGPT	LLaMA	ChatGPT	LLaMA	ChatGPT	LLaMA
Vanilla		0.58/0.48	0.44/0.39	0.62/0.55	0.57/0.56	0.29/0.42	0.25/0.37	0.53/0.46	0.56/0.46
CoT		0.46/0.51	0.56/0.42	0.49/0.54	0.51/0.55	0.33/0.51	0.35/0.52	0.73/0.62	0.59/0.51
QKR		0.58/0.53	0.44/0.39	0.64/0.57	0.62/0.60	0.58/0.46	0.64/0.47	0.39/0.51	0.43/0.47
ours	@0.1	0.72/0.84	0.77/0.8	**0.82**/0.84	0.75/**0.79**	0.71/0.83	0.71/0.72	0.68/0.8	0.78/0.71
	@0.5	**0.85**/0.86	**0.81/0.83**	0.7/**0.85**	0.64/0.75	**0.84**/0.81	**0.86/0.84**	**0.74/0.8**	**0.8/0.72**
	@1	0.79/**0.88**	0.72/0.76	0.77/0.84	**0.77**/0.72	0.8/**0.86**	0.78/0.82	0.7/0.82	0.73/0.63

model's understanding of questions changes when injecting knowledge of varying volume and relevance. To achieve this, we controlled the similarity confidence between nodes and input text, ranging from 40% to 90%, resulting in 10, 50, and 100 additional knowledge nodes as controls obtained from the results of entity retrieval. In presenting the experimental results, the three scales of node quantities are displayed as @0.1, @0.5, and @1.

VI. RESULTS

As shown in Table 2, our method consistently outperforms the comparison experiments under various conditions.

- **We successfully alleviated the model's illusions in general Q&A scenarios by employing entity enhancement methods based on KGCN retrieval and candidate screening.** On the three Q&A datasets, our proposed framework shows significant improvements compared to the Vanilla baseline. This indicates that the approach of combining knowledge retrieval and injection can effectively correct the hallucinations of LLM. Furthermore, when compared to the QKR method, which also uses a certain degree of knowledge retrieval, our method achieves better evaluation metrics. This highlights the effectiveness of the Candidate Verification Module.
- **In the context of specific complaint scenarios, we have achieved a relatively low-error, user-friendly question and answer interaction.** Compared to three baseline methods, our processing pipeline yielded the best results across all four complaint topics. This indicates the effectiveness of our framework in addressing the target issues.
- **In comparative experiments involving varying volumes and relevance of knowledge, we observed that an excess of additional knowledge can adversely impact model accuracy.** For the limited processing capacity of LLM, we found that more relevant knowledge does not necessarily translate to better performance; instead, the key lies in the strength of relevance. This is because LLM tends to focus on vocabulary related to the given question, and an abundance of knowledge can dilute this "specificity," leading LLM to incorrectly use less relevant or even erroneous entity relationships in the limited output text.

In summary, our approach consistently outperforms other methods under various conditions, demonstrating robust generalization capabilities. The findings indicate that our framework is more reliable. Furthermore, the results highlight the significance of knowledge retrieval for enhancing LLM's specialization in specific domains, where the challenge lies in constructing smaller sets of entities with higher relevance.

VII. ERROR ANALYSIS

In all the experimental results obtained above, the ChatGPT and LLaMA enhanced by our knowledge retrieval and validation framework still exhibited noteworthy errors in certain questions. We conducted a manual analysis of 150 randomly selected error samples to identify potential causes:

Authorized licensed use limited to: Vignana Bharathi Institute Of Technology. Downloaded on August 05,2025 at 06:02:09 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
DISTRIBUTION OF RANDOM 150 ERROR SAMPLES ON EACH DATASET.

Error type	Simple Question	Mintaka	HotpotQA	GCD
Retrieval Error	**96**	39	47	28
Inference Error	10	**104**	**89**	**76**
Factual Error	9	0	5	3
Other Error	35	7	9	43

Retrieval Error Utilizing the KGCN module for relevant entity retrieval resulted in obtaining incorrect entity information, leading to errors.

Inference Error Although the KGCN module provided accurate relevant knowledge, the LLM engaged in faulty reasoning during the text generation phase, and the candidate verification mechanism did not entirely eliminate such errors.

Factual Error Errors of a factual nature occurred due to the inability to find entirely relevant knowledge in the knowledge graph.

Other Error These encompassed issues such as low-quality generated text and failure to faithfully adhere to the requirements of prompt words.

In the Q&A datasets and the GDC dataset, the distribution of error types varies. Errors in the Q&A datasets are primarily attributed to Retrieval Error and Inference Error, while the GDC dataset mainly exhibits Inference Error. This discrepancy arises from the specialized design of the KG-Complaint used in the experiments, tailored for complaint scenarios. Even though we incorporated a module in which the LLM model rejects answering questions it lacks confidence in, biases in entity prediction still occur in the Q&A datasets due to the confusion between domain-specific knowledge and general knowledge. As questions become semantically more complex, errors in Mintaka and HotpotQA tend to concentrate on Inference Error compared to the simplest Simple Question. This is because intricate questions require longer reasoning paths and consideration of more entity relationships.

On the GDC dataset, Inference Error is predominant, stemming from the complexity introduced when processing complaint texts into subgraphs and expanding them. The relationships between entities become highly intricate after augmentation, leading LLM to face challenges in reasoning with complex problems and large datasets.

To address the difficulties in LLM's reasoning when faced with complex problems, a theoretically effective approach is to enhance the semantic understanding of questions. This could involve using a separate model to decompose complex questions into several simpler sub-questions, inputting them into the model individually. Each sub-model would then match its relevant knowledge, refining and enhancing the overall reasoning process.

VIII. CONCLUSION

In this work, we propose an exceptionally effective retrieval enhancement scheme that significantly reduces the occurrence of illusions in human-machine communication under the context of complaint scenarios for LLMs. Faced with an interactive request involving a complaint text, the model first transforms the input complaint text into a graph representation and evaluates it using a pre-trained KGCN model. This process generates a set of accurate multi-labels and associated nodes through a semi-random and hierarchical expansion approach. Subsequently, employing semantic consistency metrics and node prediction confidence as criteria, we filter and extract crucial additional knowledge. These relevant data are embedded into a carefully designed prompt, which is then input together with the original text into the model. Finally, the LLM outputs a deconstructed normative result based on the prompt's requirements.

On our multi-topic dataset, the proposed entity-enhancement module, after being bridged with various LLMs, consistently produces better text generation results compared to previous approaches.Our entity-enhancement module, centered around the main concept of knowledge extraction, also demonstrates outstanding performance in Question-Answering tasks on open datasets. This underscores the effectiveness of our knowledge extraction approach as the most valuable method.

IX. LIMITATIONS

Despite delivering a favorable interactive experience in mitigating model illusions through the incorporation of the entity-enhancement module, the Complaint-LLM faces constraints in its practical application and productivity within government departments, primarily due to limitations in online operational speed and response time. On an Nvidia Tesla V100, the architecture exhibits an average latency of 37 seconds per complaint text (averaging 100 words) from a dataset of 137 texts. This delay remains a formidable challenge in tasks requiring timely human-machine communication.

The primary contributor to this issue is the computational burden imposed by the extensive external knowledge. In our model, nodes in the graph generated from input text undergo tens of thousands of node similarity comparisons to ensure the accuracy of entity generation. Apart from the costly solution of hardware upgrades, a potential approach involves leveraging the concept of ensemble learning. This entails training smaller language models in different professional domains or contexts (analogous to a roundtable meeting composed of multiple models). For a given problem request, these models provide diverse perspectives, and the generated results are then constrained by a referee model or evaluation rules. This acts as the output for the entire process, offering a more efficient alternative to address the challenges associated with computational load and expedite the human-machine communication process in real-time applications.

Authorized licensed use limited to: Vignana Bharathi Institute Of Technology. Downloaded on August 05,2025 at 06:02:09 UTC from IEEE Xplore. Restrictions apply.

Acknowledgments

This research was supported by Natural Science Foundation of Shandong Province (Project No.ZR2021MF006). The authors of this article would like to especially thank the Data Annotation Team (Guangxuan Xu, Binyang Li, Xu Ren, Jia Zhang, Baoyu Li, Fangyu Liu, Jiacheng Li) at the Teddy Intelligence Studio of Shandong Jianzhu University for their hard work in building the knowledge graph. Special thanks are also due to Professor Niu for her insightful comments and feedback on earlier drafts of this paper.

References

[1] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. http://jmlr.org/papers/v21/20-074.html

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, "Language Models are Few-Shot Learners," arXiv preprint, 2020. https://arxiv.org/abs/2005.14165

[3] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, Guillaume Lample, "LLaMA: Open and Efficient Foundation Language Models," arXiv preprint, 2023. https://arxiv.org/abs/2302.13971

[4] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, "Gemini: A Family of Highly Capable Multimodal Models," arXiv preprint, 2023. https://arxiv.org/abs/2312.11805

[5] Zhonghua Zheng, Lizi Liao, Yang Deng, Liqiang Nie, "Building Emotional Support Chatbots in the Era of LLMs," arXiv preprint, 2023. https://arxiv.org/abs/2308.11584

[6] Kaize Shi, Xueyao Sun, Dingxian Wang, Yinlin Fu, Guandong Xu, Qing Li, "LLaMA-E: Empowering E-commerce Authoring with Multi-Aspect Instruction Following," arXiv preprint, 2023. https://arxiv.org/abs/2308.04913

[7] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint, 2023. https://arxiv.org/abs/2312.10997

[8] Anton Shapkin, Denis Litvinov, Timofey Bryksin, "Entity-Augmented Code Generation," arXiv preprint, 2023. https://arxiv.org/abs/2312.08976

[9] Abdelkarim El-Hajjami, Nicolas Fafin, Camille Salinesi, "Which AI Technique Is Better to Classify Requirements? An Experiment with SVM, LSTM, and ChatGPT," arXiv preprint, 2023. https://arxiv.org/abs/2311.11547

[10] Jinfeng Zhong, Elsa Negre, "Context-aware explainable recommendations over knowledge graphs," arXiv preprint, 2023. https://arxiv.org/abs/2310.16141

[11] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, Ji-Rong Wen, "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models," arXiv preprint, 2023. https://arxiv.org/abs/2305.11747

[12] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, Siva Reddy, "On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?," arXiv preprint, 2022. https://arxiv.org/abs/2204.07931

[13] Pouya Pezeshkpour, "Measuring and Modifying Factual Knowledge in Large Language Models," arXiv preprint, 2023. https://arxiv.org/abs/2306.06264

[14] Guido Zuccon, Bevan Koopman, Razia Shaik, "ChatGPT Hallucinates when Attributing Answers," arXiv preprint, 2023. https://arxiv.org/abs/2309.09401

[15] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, Noah A. Smith, "How Language Model Hallucinations Can Snowball," arXiv preprint, 2023. https://arxiv.org/abs/2305.13534

[16] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, Jianfeng Gao, "Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback," arXiv preprint, 2023. https://arxiv.org/abs/2302.12813

[17] Chao Feng, Xinyu Zhang, Zichu Fei, "Knowledge Solver: Teaching LLMs to Search for Domain Knowledge from Knowledge Graphs," arXiv preprint, 2023. https://arxiv.org/abs/2309.03118

[18] Lang Cao, "Learn to Refuse: Making Large Language Models More Controllable and Reliable through Knowledge Scope Limitation and Refusal Mechanism," arXiv preprint, 2023. https://arxiv.org/abs/2311.01041

[19] Hanseok Oh, Haebin Shin, Miyoung Ko, Hyunji Lee, Minjoon Seo, "KTRL+F: Knowledge-Augmented In-Document Search," arXiv preprint, 2023. https://arxiv.org/abs/2311.08329

[20] Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, Le Sun, "Mitigating Large Language Model Hallucinations via Autonomous Knowledge Graph-based Retrofitting," arXiv preprint, 2023. https://arxiv.org/abs/2311.13314

[21] Rico Sennrich, Jannis Vamvas, Alireza Mohammadshahi, "Mitigating Hallucinations and Off-target Machine Translation with Source-Contrastive and Language-Contrastive Decoding," arXiv preprint, 2023. https://arxiv.org/abs/2309.07098

[22] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, Scott Wen-tau Yih, "Trusting Your Evidence: Hallucinate Less with Context-aware Decoding," arXiv preprint, 2023. https://arxiv.org/abs/2305.14739

[23] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, Jason Weston, "Chain-of-Verification Reduces Hallucination in Large Language Models," arXiv preprint, 2023. https://arxiv.org/abs/2309.11495

[24] Bordes, A.; Usunier, N.; Chopra, S.; and Weston, J. 2015. Large-scale simple question answering with memory networks. arXiv preprint arXiv:1506.02075.

[25] Sen, P.; Aji, A. F.; and Saffari, A. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. arXiv preprint arXiv:2210.01613.

[26] Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. arXiv preprint arXiv:1809.09600.

[27] OpenAI. 2022. Introducing ChatGPT

[28] Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35: 2482424837.

[29] Baek, J.; Aji, A. F.; and Saffari, A. 2023. Knowledge Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. arXiv preprint arXiv:2306.04136.

[30] Jiawei Chen, Hongyu Lin, Xianpei Han, Le Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," arXiv preprint, 2023. https://arxiv.org/abs/2309.01431

Authorized licensed use limited to: Vignana Bharathi Institute Of Technology. Downloaded on August 05,2025 at 06:02:09 UTC from IEEE Xplore. Restrictions apply.