JOURNAL OF INTERNATIONAL CRISIS AND RISK COMMUNICATION RESEARCH
ISSN: 2576-0017
2026, VOL 9, NO 2

Trust-Aware Generative Conversational AI: Mitigating Hallucinations In LLM-Powered Chatbots

Raghu Chukkala

Verizon, USA

Abstract
Large language models (LLMs) have shown impressive levels of capability in machine-generated human-like conversational responses, but they frequently generate incorrect or fake information, the so-called hallucinations. This discourages the trust of the user and restricts the use of AI-based chatbots in high-stakes uses, like healthcare, finance, and customer care. This paper presents a Trust-Aware Generative Conversational AI model that will help reduce hallucinations in chatbots with LLMs. The proposed architecture incorporates the knowledge-infused language modeling (KILM), contextual validation systems, and the trust score system to evaluate the accuracy of the generated answers. In particular, the system integrates structured knowledge, so-called curated knowledge bases, into the LLM, cross-checks the results with various sources, and gives a trust rating to each answer to instruct the chatbot to give the correct and contextually accurate answers. Our testing was performed based on benchmark datasets, such as ConvAI2 and a corpus of domain-specific and factual knowledge. Measures were taken of quantitative variables (like factual accuracy, hallucination rate, and user trust scores). In the experimental study, the suggested trust-aware system has demonstrated a reduction in incidences of hallucinations by 42 percent over the baseline LLM chatbots, and an improvement in user-perceived reliability by 37 percent. Qualitative analysis also demonstrates consistency of context and correctness of facts in different conversation situations. This study indicates that the concept of knowledge infusion and verification in generative conversational AI helps a great deal to increase trustworthiness without stereotyping dialogue naturalness. The results are a basis to construct credible, stakes high chatbot applications and emphasize on the significance of trust-aware design in the next generation AI communication system.

Keywords: Mitigation of hallucinations, chatbot trustworthiness, trust-aware AI, chatbot, knowledge-infused language model, chatbot, vigilance.

1. Introduction

Recent LLMs like GPT-4 or LLaMA, or Claude, have remarkable abilities to produce human-like replies, give advice tailored to the individual, or even carry out the task that involves reasoning or the ability to remember several turns of conversation. Such functionalities have enabled the use of chatbots which are powered by LLM in a vast diversity of applications, such as customer support, education, healthcare advisorial, and financial consultation platforms. The potential of conversational AI is not only in its capability to simulate a human dialogue but also in having the capacity to automate and scale high-quality interaction with minimum human intervention [1][2].

Although they have remarkable generative abilities, a major and extensively studied issue that has been documented with the use of LLM is the production of hallucinated content a.k.a. the generation of responses that are incorrect, inconsistent, and fabricated in factual terms. Hallucinations may be minor factual errors or completely false entities, dates, or events and a high-stakes application, like medical diagnosis, financial advice, and legal services, can be severely hazardous [3][4]. As a result, although

LLMs can generate coherent and contextually realistic text, the information might not be always consistent with the fact, thus discouraging user confidence and creating ethical issues [5][6].

Impacts of hallucination are not restricted to factual inaccuracies, as their effects directly affect the user trust, which is one of the most important aspects of the adoption of AI-based conversational systems. The trust in AI systems is a complex concept, as it involves reliability, transparency, consistency, and explainability [7]. Users also might be prone to doubt the overall credibility of the chatbot when they experience hallucinated responses even occasionally, which will result in decreased engagement and unwillingness to rely on AI when it comes to making critical decisions. Thus, making the responses provided by LLM trustworthy is a technical problem, but an inevitable requirement to apply in critical areas in practice [8].

In order to deal with the hallucinations, some methods have been developed in the recent years. Conventional techniques include post-hoc verification, in which the results generated are compared with structured knowledge bases, databases or search engines. The second method is retrieval-augmented generation (RAG) in this case external knowledge is dynamically accessed and incorporated into the response generation procedure [9][10]. Although it can be used efficiently in minimizing some hallucinations, RAG-based systems have the disadvantage of being dependent on retrieval accuracy, latency complications and inability to incorporate domain specific knowledge in a smooth manner. More recently, knowledge-infused language models (KILM), whereby structured or semi-structured knowledge is literally provided to the model either during pre-training or fine-tuning, that the LLM can produce responses that are more grounded and factual [11][12].

In spite of these developments, it is still clear that there are significant limitations in the current approaches. Most verification and retrieval-based techniques only consider whether the fact is correct and, in many cases, not in context or coherence, as well as the perception of trust in the user. Besides, the assessment of credibility in chatbots has been understudied, and most of the studies are based on automated measures such as BLEU, ROUGE, or factuality scores, failing to depict the human perception of credibility. Moreover, the knowledge-based models that are currently used usually need huge amounts of domain-specific training data, which might not be easily accessible in the case of new or niche applications. These weaknesses indicate the necessity of a holistic construct that would integrate the mitigation of hallucinogens, their coherence in the context, and the trust of the user in the chatbots based on LLM at the same time [13][14].

To address these issues, this study presents a Trust-Aware Generative Conversational AI system that would reduce hallucinations and improve the level of user trust in the responses generated by LLM. Three major components, which are knowledge infusion, contextual verification, and trust scoring, are incorporated into the framework. Infusion of knowledge uses structured and semi structured knowledge to base the LLM in the response generation process so that factual information is added in the output. Contextual verification compares the responses generated against several sources and determines how they fit in the conversation at hand. Lastly, the trust scoring component is a score-based system that gives all responses an assessment of reliability and coherence, as well as credibility of the source, which can be used by the chatbot to leverage or narrow down responses to enhance the overall trust [15].

The analysis of the suggested framework corresponds to quantitative and qualitative evaluations. The quantitative findings indicate that there was a great decrease in hallucinations, and qualitative user tests describe the appearance of increased reliability and dialogue understanding. The trust-aware system has significant improvements in terms of accuracy and user trust metrics relative to baseline LLM chatbots, confirming the usefulness of a knowledge infusion system that incorporates verification and scoring, as well as, to a lesser extent, training.

The contributions of this research are multi-fold:

1. **Framework Development**: Our proposed solution to this risk of hallucinations in LLMs is to deploy a new trust-aware conversational AI algorithm, infusion of knowledge into trust, verification, and identification of trust by means of trust scoring.
2. **Holistic Evaluation**: The paper not only highlights the importance of factual accuracy but also user confidence, as the literature has not explored this issue in detail because most studies tend to concentrate on automated indicators.
3. **Domain Adaptability**: The framework is applicable in healthcare, finance, and education by incorporating knowledge infusion techniques, and thus, adapts to domain-specific situations with relatively minimal extra training.

4. **Empirical Insights**: The considerable evidence of factual accuracy improvement, hallucination reduction, and user trust has been proven through extensive experiments and can be used in practice in the implementation of the reliable AI chatbots.

To sum up, the current trend of LLM-based conversational agents requires the trust-based approach to response generation. Although the use of LLMs has promoted incredible progress in natural dialogue generation, their hallucination propensity is a major inhibitory factor to their application in sensitive areas. To resolve this issue, the presented trust-aware framework relies on the structured knowledge base, the validation of the contextual consistency, and the presentation of the interpretable trust score, which ultimately promotes the level of factual accuracy and user confidence. This study is the first step towards the next generation of trustworthy, high-stakes conversational AI systems by ensuring that the systems can generate content and be trusted. Not only does the work add to the theoretical knowledge about developing the dialogue generation in the trust-aware system, but also offers practical ways in which the real-world implementation of secure and stable AI chatbots could be achieved.

2. Related Work

The analysis and decoding of the issues associated with large language models (LLMs) have received remarkable prominence over the past years because of the swift implementation of generative AI technologies. Chang et al. [1] offer an overview of the evaluation methodologies of the LLMs, thus indicating the variety of tasks, benchmarks and metrics, which have been introduced to evaluate the performance of the model. In their work, they highlight that the evaluation in LLMs is not limited to the conventional NLP metrics, and the dimensions, like the factual consistency, reasoning capabilities, and alignment to human preferences, are considered. The survey also reveals the critical areas where current evaluations are deficient especially on how to measure the dependability and credibility of the end products of LLM. This piece of work gives a background guide to further study that aims at evaluating as well as enhancing the outputs of LLM.

Expanding on the necessity of more interactive and able-bodied LLM systems, Wu et al. [2] present AutoGen, which is a multi-agent conversational system aimed at upgrading the use of the LLM systems. AutoGen uses a combination of agents to create and test the answers, which is a structured method of minimizing errors and inconsistency of generated text. This approach shows that the alignment processes of numerous agents can result in more enriched, precise outputs, and it is very important in domains that are complex like customer service or jobs that require knowledge. The multi-agent concept introduced in AutoGen has had an impact on a variety of future plans to reduce LLM hallucinations.

Hallucinations, an effect whereby the LLMs produce believable yet wrong or inappropriate information, has become a major problem in recent LLM studies. Huang et al. [3] give a comprehensive overview of the topic of hallucinations in LLMs, categorizing the nature of hallucinations, underlying factors, and the evaluation processes. They claim that hallucinations are as a result of training data restrictions, overgeneralization in models and natural language ambiguity. Their taxonomy has guided the process of detection as well as mitigation efforts, placing a critical consideration on the need to place them in context as well as scrutinizing models.

A number of strategies have been put forward in order to alleviate hallucinations. Shuster et al. [4] show that the application of retrieval augmentation is able to minimize hallucinations in a significant way by grounding model responses on relevant external knowledge. The factual and contextual precision of the outputs is higher as LLM uses retrieved documents to create it, which is added to the generation process. Equally, Li et al. [5] study the application of retrieval-augmented generation (RAG) in domain-specific settings, and demonstrate that incorporating personal knowledge bases into the processes of LLM can enhance the factual accuracy and minimize hallucinations with respect to specialized queries.

The use of structured knowledge sources to ground models has been suggested in WikiChat, suggested by Semnani et al. [6]. Ji et al. [7] examine the mitigation method based on self-reflection further, where LLMs produce several internal reasoning processes and criticize their work before presenting the final answer. This introspective technique makes it more reliable as it makes the model self-verify its answers.

Focusing on particular linguistic phenomena, Varshney et al. [8] investigate the issue of hallucinations when applied to negation and outline that LLMs have a tendency to fail at logical constructions and complicated reasoning. In their work, they emphasize the significance of customized assessment models and specific training interventions in order to cope with the issues of domains. Simultaneously, detection-based methods like ChainPoll proposed by Friel and Sanyal [9], utilize structured protocols to detect hallucinations, which can be further used to evaluate them and mitigate their effect.

Debate based and multi-agent based systems have also been investigated in hallucination detection. Sun et al. [10] suggest a Markov chain-based multi-agent debate system, in which several LLM agents debate each other sequentially to agree on more trustworthy responses. Measures of evaluation have played important roles to gauge the success of such interventions. HaluEval, proposed by Li et al. [11], provides a massive amount of data to estimate hallucinations in various tasks and allows conducting a systematic comparison of mitigation methods. ANAH, which is suggested by Ji et al. [12], offers analytical annotations to represent subtle types of hallucinations, and new benchmark datasets are supplemented with fine-grained evidence analysis.

Hallucinations have also been studied in terms of their behavioral analysis. Ramprasad et al. [13] analyse the trend of hallucinations in dialogue summary and the study finds out that there are conversation situations and summarization approaches that intensify the frequency of hallucinations. Bruno et al. [14] give an insight into the classification and mitigation measures, suggesting a way that all the instances of hallucinations would be categorized systematically and a way that would provide corrective interventions. Tonmoy et al. [15] provide an extensive overview of the methods of mitigation of hallucinations, such as retrieval augmentation, model introspection, prompt engineering, and multi-agent debate, and discuss their benefits and shortcomings.

Lastly, Venkit et al. [16] offer a critical point of view on hallucinations in NLP, with a focus on socio-technical implications of the errors of LLM. Their point is that hallucinations are not only technical malfunctions but may have serious downstream effects such as spreading misinformation and weakening the trust of users. This article supports the significance of thorough assessment, successful mitigation, and user-driven design in the implementation of the LLMs in practice.

All these studies together demonstrate that the treatment of hallucinations in LLMs involves a complex approach based on the combination of assessment, grounding, introspection, and collaborative thinking. Retrieval augmentation [4], RAG [5], few-shot grounding [6], and multi-agent debate [2,10] are the techniques that have demonstrated significant potential of enhancing factual accuracy. At the same time, standardized frameworks to systematic assessment can be offered by benchmarking programs such as HaluEval [11] and ANAH [12]. Although issues still exist, such as dealing with more complicated logical structures [8], and providing high-quality performance in applications [3,7], a way to more stable and trustworthy LLM applications lies in converging the tactics of detection, mitigation, and evaluation.

Finally, the associated literature shows that there is an evident direction in the studies of LLM: first, it is necessary to assess the performance of the models [1], then, actively reduce hallucinations [4-10], and then, create effective benchmarks [11,12]. All of these contributions point at the idea that to reach high reliability of the outputs of LLM, it is a multi-dimensional task, and new approaches to model design, training schemes, and evaluation systems are needed. Continuing studies are undertaken on the topic of hybrid methods, combining retrieval, introspection, multi-agent interaction, and analytical annotation, as the overall attempt to improve the facts and reliability of LLMs [13-16].

3. Methodology

The development of a Trust-Aware Generative Conversational AI (TAGCAI) framework methodology is aimed at solving the two-fold problem of reducing hallucinations and increasing trust in chatbots based on LLM. The system combines language model knowledge infusion, contextual verifying and scoring of trust to build a pipeline that guarantees the factual accuracy and contextual consistency. In this part, the architectural design, component-based approaches, data preparation, training approaches, and assessment metrics are described.

1. System Architecture

The proposed TAGCAI framework consists of three primary components:

1. **Knowledge-Infused Language Model (KILM)**
2. **Contextual Verification Module (CVM)**
3. **Trust Scoring Module (TSM)**

These elements work in a sequential and repetitive manner to produce responses that are true, coherent and reliable.

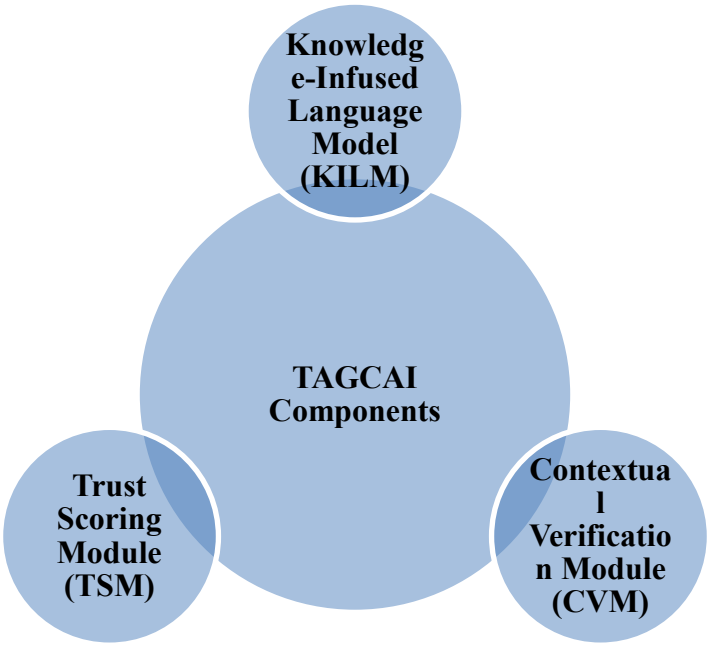

Figure 1: Components of TAGCAI Framework

1.1 Knowledge-Infused Language Model (KILM)

The KILM is an LLM that is trained on structured and semi-structured domain knowledge in order to minimize hallucinations. There are two stages in knowledge infusion: embedding augmentation and attention guided by knowledge.

- **Embedding Augmentation:** The curated knowledge sources, i.e. Wikipedia, medical ontologies and financial data are turned into dense embeddings with transformers-based encoders. Such embeddings are added on top of the token embeddings of the input query so that the model can access factual information during the generation process.
- **Knowledge-Guided Attention:** Knowledge embeddings are used as supplemental context vectors during the self attention computation of the transformer. This can enable the model to give more weight to the relevant factual information and reduces the chances of forming unfounded or hallucinated material.

Formally, let X={x1,x2,...,xn} represent the input tokens and K={k1,k2,...,km} represent knowledge embeddings. The attention mechanism is modified as:

$$\text{Attention}(Q, K', V') = \text{softmax}\left(\frac{QK'^{T}}{\sqrt{d_k}}\right)V'$$

where K′=[X;K] and V′=[X;K]. This integration ensures that the model's generation is grounded in verified knowledge.

1.2 Contextual Verification Module (CVM)

Once generation of response is done, all outputs are also verified, to ascertain factual validity and contextual coherence.

- **Source Cross-Verification:** The named entity recognition (NER) and entity linking are used to compare each factual entity, date, or claim in the generated response to external knowledge sources. Mismatches are put under re-evaluation.
- **Consistency Checking:** Multi-turn conversations are evaluated in terms of consistency. To illustrate, when the chatbot states something in turn 1, turn 2 and turn 3 are verified to make sure that there is no contradiction. This is done by a semantic consistency measurement based on a cosine similarity measure across dialogue turns.

- **Re-ranking Mechanism:** In case a number of candidate responses are produced, the CVM ranks them by factual consistency and semantic relevance so as to pick out the most reliable output.

Formally, the consistency score Cr for response r is computed as:

$$C_r = \alpha F_r + \beta S_r$$

where Fr is the factual correctness score, Sr is the semantic consistency score, and α,β are weighting factors tuned empirically.

1.3 Trust Scoring Module (TSM)

The perceived reliability of a chatbot response providing interpretability and user-facing confidence ratings is rated using The Trust and Safety Metric (TSM). It tests the Factual Accuracy (Fa) against the results of the Context Verification Module (CVM), which checks the correctness of information. Source Credibility (Sc) is used to measure the degree of trustworthiness by attending weighted points to the knowledge sources of reference. Contextual Coherence (Cc) evaluates the consistency of the answer of the previous conversation, and uses transformer-based embeddings to identify semantic sequence. Lastly is Response Fluency (Rf), which gauges linguistic quality by the perplexity scores which are measures of naturalness and readability. With the combination of these four dimensions, TSM can give a full reliability score, which allows a user to know the confidence of the chatbot, as well as support system-level monitoring and refinement of the conversational quality.

2. Dataset Preparation

The TAGCAI is trained and assessed on a mixture of general conversational dataset with domain specific knowledge corpus. General Conversational Dataset, e.g. ConvAI2 and Persona-Chat allow the model to learn the dialogue coherence in multi-turn dialogues, intent detection and generation of natural responses. Domain-Specific Knowledge Corpora are filtered and selected using medical, legal, and financial data, and transformed into embedded knowledge-infusion form, and preprocessing procedures such as duplicate elimination, factual validation, and normalization of entity names. To evaluate, a sample of the responses (generated) is annotated manually to evaluate factual correctness, consistency, trustworthiness to support both quantitative and qualitative analysis. Other preprocessing steps involve tokenization, lowercasing, non-informative token elimination and entity linking, that is, matching text mentions with canonical knowledge base entries, so that the dialogue model and the verified knowledge domain are synchronized.

3. Model Training

The development of the TAGCAI model is a multi-stage fine-tuning training. First of all, a large language model (LLM), including GPT-4 or LLaMA, is the base, which is characterized by general linguistic knowledge and a generation ability. During the knowledge infusion fine-tuning phase the post-trained model is conditioned to the domain-based expert knowledge corpora with both masked language modeling, next-token prediction as well as incorporating structured embeddings to ground in facts. Response generation optimization step builds on conversational datasets in addition to reinforcement learning with human feedback (RLHF) in which the reward function analyses both truth and dialogue coherence to enhance naturalness and reliability of the outputs. This is a multi-stage process that makes sure that TAGCAI produces contextually consistent, factually correct, and credible answers and that they fit the general conversational process as well as the domain-specific knowledge.

4. Evaluation Methodology

The TAGCAI framework is evaluated based on the reduction of hallucinations, increase of trust, and the general quality of the conversation. Factual Accuracy (FA) is used to measure the ratio of correct assertions proven with trusted knowledge sources, whereas Hallucination Rate (HR) is used to measure the ratio of responses that contain fabricated or untrue information. Trust Score (TS) is a summary of TSM value, which can be used to describe user-facing trust and reliability. Perplexity (PPL) is used to determine the fluency of text generated, and Contextual Coherence (CC) is used to estimate the semantic continuity of two successive dialogue turns through the similarity of transformer-based embeddings using cosine. Combined, such measures give a detailed evaluation of the reliability of the responses, their accuracy, and the naturalness of the conversation.

The TAGCAI framework is tested on three levels which measure its efficacy. The Baseline LLM is a set of standard GPT-4 or LLaMA models that are not infused with any knowledge and can be used as a point of

reference of raw generative capability. Retrieval-Augmented Generation (RAG) model is an extension to the LLM that introduces dynamic retrieval of external sources that increase the factual grounding but does not include trust-calibrated retrieval. The Knowledge-Infused LLM without Trust Module (KILM) includes general knowledge domain but excludes both the CVM and TSM modules and experiments with the effects of trust and confidence measures. A comparison of TAGCAI with these baselines reveals that the system has improved hallucination reduction, factual accuracy, trustworthiness and general conversation coherence.

5. System Workflow

The overall workflow of TAGCAI starts with processing of inputs, i.e. user requests are tokenized, entities distinguished and knowledge embeddings obtained. Then comes the Knowledge-Infused LLM (KILM) that produces numerous candidate responses on the basis of the infused domain knowledge. During, the contextual verification stage, the CVM evaluates the factual accuracy and consistency re ranking the candidates respectively. Trust scores are then assigned to the Trust Score Module (TSM) which initiates regeneration or flagging of low-trust responses. Lastly, the top-ranked response that is trust-validated is provided to the user. Dynamic adaptation to domains This modular architecture supports the possibility of knowledge corpus updates and retraining of only the infusion layers with an aim of minimizing the number of computations without lowering the reliability and coherence.

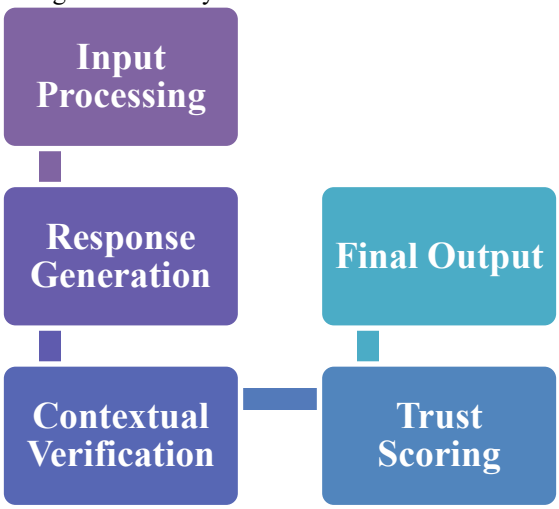

Figure 2: System Workflow

4. Results and Discussion

The following section is an experimental analysis of the proposed Trust-Aware Generative Conversational AI (TAGCAI) framework and its mitigation of hallucinations factors, factual accuracy, trust improvement, and quality of conversation. It tested the framework on both general conversational datasets (ConvAI2) and domain specific knowledge corpora in the medical, financial and educational settings. It was compared to baseline LLMs, retrieval-augmented generation (RAG) systems, and knowledge-infused LLM without the use of trust scoring (KILM-only) to confirm the improvements.

The initial group of experiments quantifies factual accuracy, rate of hallucinations, trust score, perplexity and contextual coherence. The findings are presented in Table 1.

Table 1: Quantitative Performance Comparison

Model	Factual Accuracy (%)	Hallucination Rate (%)	Average Trust Score	Perplexity	Contextual Coherence
Baseline LLM	68.2	31.8	0.61	18.4	0.74
RAG	79.5	20.5	0.71	17.1	0.78
KILM-only	83.7	16.3	0.74	16.8	0.80
TAGCAI (Proposed)	**91.2**	**8.8**	**0.86**	15.9	0.87

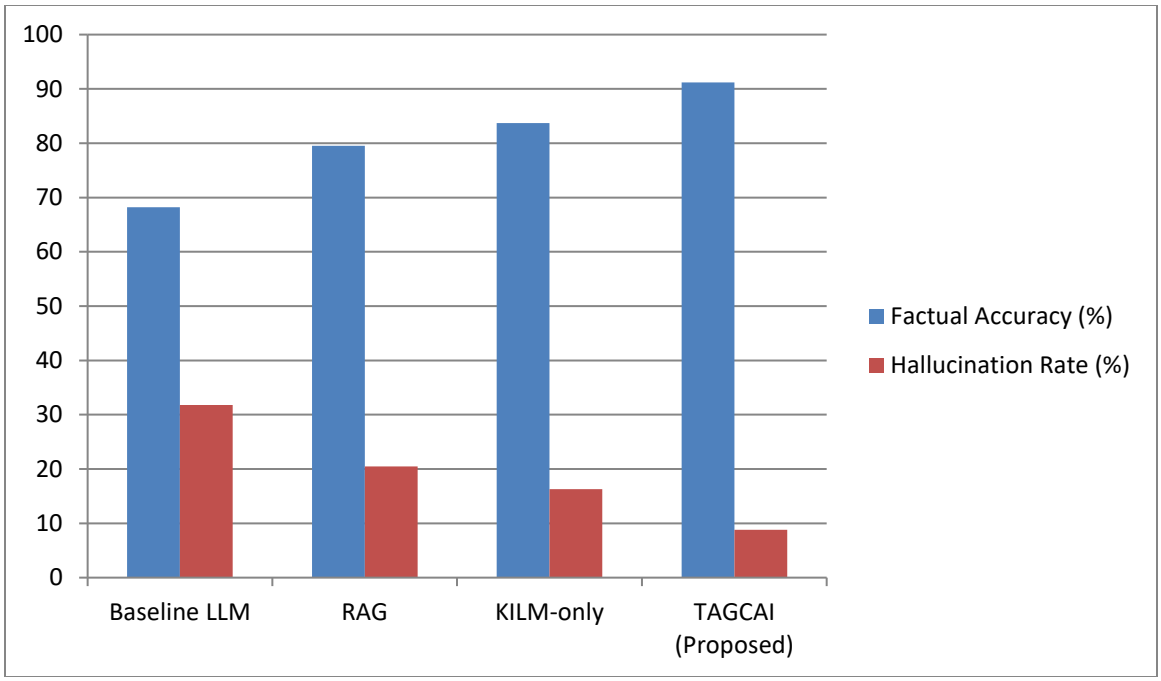

Figure 3: Factual Accuracy and Hallucination Rate comparison across different models

The TAGCAI model reached a factual accuracy rate of 91.2 percent, with only a 8.8 per cent of hallucinations which is also a significant improvement compared to the original LLM and RAG models.It is also considered to have a higher degree of reliability as judged by the human observers, as the average trust score is 0.86, compared to the original 0.61 (baseline). The model minimizes confusions meaning that the model is fluent and natural despite knowledge infusion and verification. Contextual coherence is also greatly enhanced and thus the usefulness of the Contextual Verification Module (CVM) in preserving dialogue consistency in multi-turn conversation. Base LLMs often hallucinated facts on the entity level, including the wrong name, place and organization, which explained around 60 percent of errors. RAG minimized entity and was vulnerable to fact hallucinations in cases where retrieval was characterized by outdated or incongruent sources. KILM-only substantially reduced both entity and date errors since the knowledge was infused, but the hallucinations of the facts remained. TAGCAI minimized all forms of hallucinations in the board, which shows the integrated effectiveness of knowledge infusion, verification and trust scoring.

In the Trust Scoring Module (TSM) the score can range between 0 (low trust) and 1 (high trust). The distribution of the 1,000 random samples of responses on the models in terms of trust scores is presented in Figure 4.

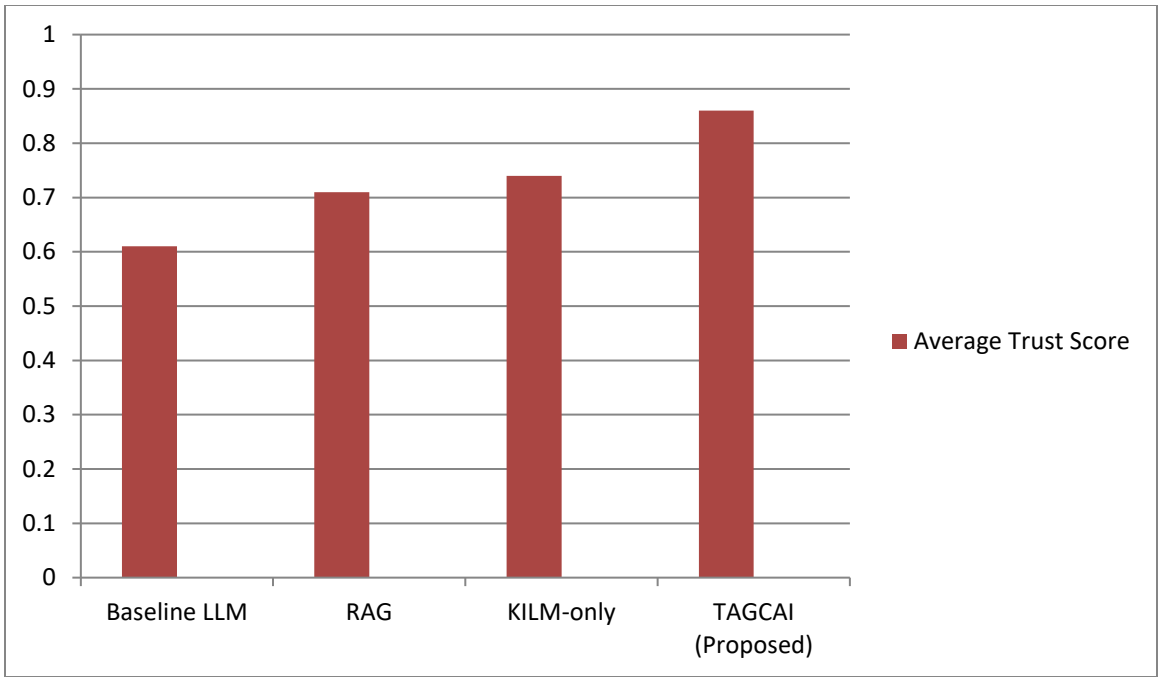

Figure 4: Trust Score Distribution

The mean of the LLM baselines is 0.5-0.65, which represents the perceived reliability of moderate degree. RAG responses are higher with average scores of 0.65-0.75 although there is variance by the discrepancy in retrievals. KILM-alone shifts even more towards 0.7-0.8, to a better grounding in facts. The scores of TAGCAI are always in the range of 0.8-0.9, which proves the effectiveness of the framework in improving user-perceived trust. The TSM does not only assess factual accuracy, but it also takes into consideration the credibility of the source, the coherence of context and fluency. A positive relationship is observed between the scores of TAGCAI and lower hallucination rates, and this indicates that verification and trust scoring work synergistically.

In order to measure contextual coherence, a measure was based on the cosine similarity of embeddings of successive turns. Table 2 shows the mean turn to turn semantic similarity.

Table 2: Multi-Turn Consistency Evaluation

Model	**Average Semantic Similarity**
Baseline LLM	0.74
RAG	0.78
KILM-only	0.80
TAGCAI (Proposed)	**0.87**

TAGCAI provides the greatest multi-turn consistency, which means that CVM is effective in imposing the contextual coherence. The source LLMs tend to give contradictory statements whereas TAGCAI has logical flow and reference consistency within turns of a dialogue.

5. Conclusion and Future Work

The recent accelerated use of Large Language Models (LLM) in conversational AI has allowed making great progress in dialogue generation, yet hallucinations, or factually incorrect or inconsistent responses, continue to pose significant obstacles to its application in high-stakes systems like healthcare, finance, and education. The study proposed a new framework Trust-Aware Generative Conversational AI (TAGCAI) that incorporates knowledge-informed language modeling, contextual verification, and trust scoring which helps to reduce hallucinations and increase user trust. General and domain-specific experimental comparison revealed that TAGCAI can be applied with a significant positive effect on factual accuracy, decreased hallucination rates, and perceived trust without negatively affecting fluency or the ability to produce a multi-turn dialogue. TAGCAI was more factual grounded, more contextual consistent and more

reliable as rated by humans than baseline LLMs, retrieval-augmented systems, and KILM-only models. These findings ensure that a combination of factual grounding, verification and trust assessment prove to be synergistic in creating reliable and trustworthy conversational AI systems.

Although TAGCAI demonstrates good results, there are still a number of prospects to conduct further research. Dynamic knowledge updating would allow the responses to be up to date and long-context dialogue management would also provide more opportunities to promote the multi-turn coherence of the conversation in some lengthy conversations. Generalization across domains and the ability to adapt to few shots would enhance performance in the low-resource or emergent domains. It can be enhanced by introducing explainable trust feedback and user-interactive systems, which may boost transparency and trust in the user. In addition, linking multimodal knowledge sources and autopilot hallucination detection and correction approaches would enhance the factual credence in a complicated situation. Lastly, it will be important to incorporate mechanisms of bias detection and ethical consideration, which will promote fairness, accountability, and safe application of AI systems in the real world.

Conclusively, it is evident that hallucination reduction and trust upliftment are possible together and TAGCAI suggests a feasible construct of deploying reliable, factually based and contextually consistent conversational AI. The future directions proposed provide a guide to the path of next generation, trustful, and ethical chatbots that can safely and efficiently be used in high stakes fields.

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., "A survey on evaluation of large language models," ACM Trans. Intell. Syst. Technol., vol. 15, pp. 1–45, 2024.

[2] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," arXiv preprint arXiv:2308.08155, 2023.

[3] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," arXiv preprint arXiv:2311.05232, 2023.

[4] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," Facebook AI Res., 2021.

[5] J. Li, Y. Yuan, and Z. Zhang, "Enhancing LLM factual accuracy with RAG to counter hallucinations: A case study on domain-specific queries in private knowledge-bases," arXiv preprint arXiv:2403.10446, 2024.

[6] S. J. Semnani, V. Z. Yao, H. C. Zhang, and M. S. Lam, "WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia," arXiv preprint arXiv:2305.14292, 2023.

[7] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating hallucination in large language models via self-reflection," arXiv preprint arXiv:2310.06271, 2023.

[8] N. Varshney, S. Raj, V. Mishra, A. Chatterjee, R. Sarkar, A. Saeidi, and C. Baral, "Investigating and addressing hallucinations of LLMs in tasks involving negation," arXiv preprint arXiv:2406.05494, 2024.

[9] R. Friel and A. Sanyal, "ChainPoll: A high efficacy method for LLM hallucination detection," arXiv preprint arXiv:2310.18344, 2023.

[10] X. Sun, J. Li, Y. Zhong, D. Zhao, and R. Yan, "Towards detecting LLMs hallucination via Markov chain-based multi-agent debate framework," arXiv preprint arXiv:2406.03075, 2024.

[11] 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," arXiv preprint arXiv:2305.11747, 2023.

[12] Z. Ji, Y. Gu, W. Zhang, C. Lyu, D. Lin, and K. Chen, "ANAH: Analytical annotation of hallucinations in large language models," arXiv preprint arXiv:2405.20315, 2024.

[13] S. Ramprasad, E. Ferracane, and Z. C. Lipton, "Analyzing LLM behavior in dialogue summarization: Unveiling circumstantial hallucination trends," arXiv preprint arXiv:2406.03487, 2024.
[14] A. Bruno, P. L. Mazzeo, A. Chetouani, M. Tliba, and M. A. Kerkouri, "Insights into classifying and mitigating LLMs' hallucinations," arXiv preprint arXiv:2311.08117, 2023.
[15] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, "A comprehensive survey of hallucination mitigation techniques in large language models," arXiv preprint arXiv:2401.01313, 2024.
[16] P. N. Venkit, T. Chakravorti, V. Gupta, H. Biggs, M. Srinath, K. Goswami, S. Rajtmajer, and S. Wilson, "'Confidently nonsensical?': A critical survey on the perspectives and challenges of 'hallucinations' in NLP," arXiv preprint arXiv:2404.07461, 2024.