

# Natural Language Understanding (CSL 7640)

## Minor 2

Submitted By – **Debonil Ghosh [M21AIE225]**

Q1. Read the paper to understand issues about “hallucinations” that are encountered in NLG tasks.

The paper "Survey of Hallucination in Natural Language Generation" provides an overview of research progress and challenges in the hallucination problem in NLG. The survey is organized into two parts: a general overview and a task-specific overview. The general overview covers metrics, mitigation methods, and future directions for addressing hallucination in NLG. The task-specific overview provides an in-depth analysis of hallucination in six downstream tasks: abstractive summarization, dialogue generation, generative question answering, data-to-text generation, machine translation, and visual-language generation. Overall, the paper aims to provide researchers with a high-level insight into the similarities and differences between different approaches to addressing hallucination in NLG.

Q2. Use Chatgpt (or any other NLG system) to generate text for two different tasks of your choice from the below-mentioned list in such a way that the system is “forced” to generate “hallucinations.” (5 marks)

Task 1 – one to be picked up from the set

{Dialogue generation, Question Answering, Abstractive summarization}

Task 2 – one to be picked up from the set

{Data-to-Text, Machine Translation (English to Indian Language and Vice-versa)}

**Answer.**

### **Task 1 – Abstractive summarization**

Prompt: "Summarize the plot of the movie 'The Shawshank Redemption' in one sentence."

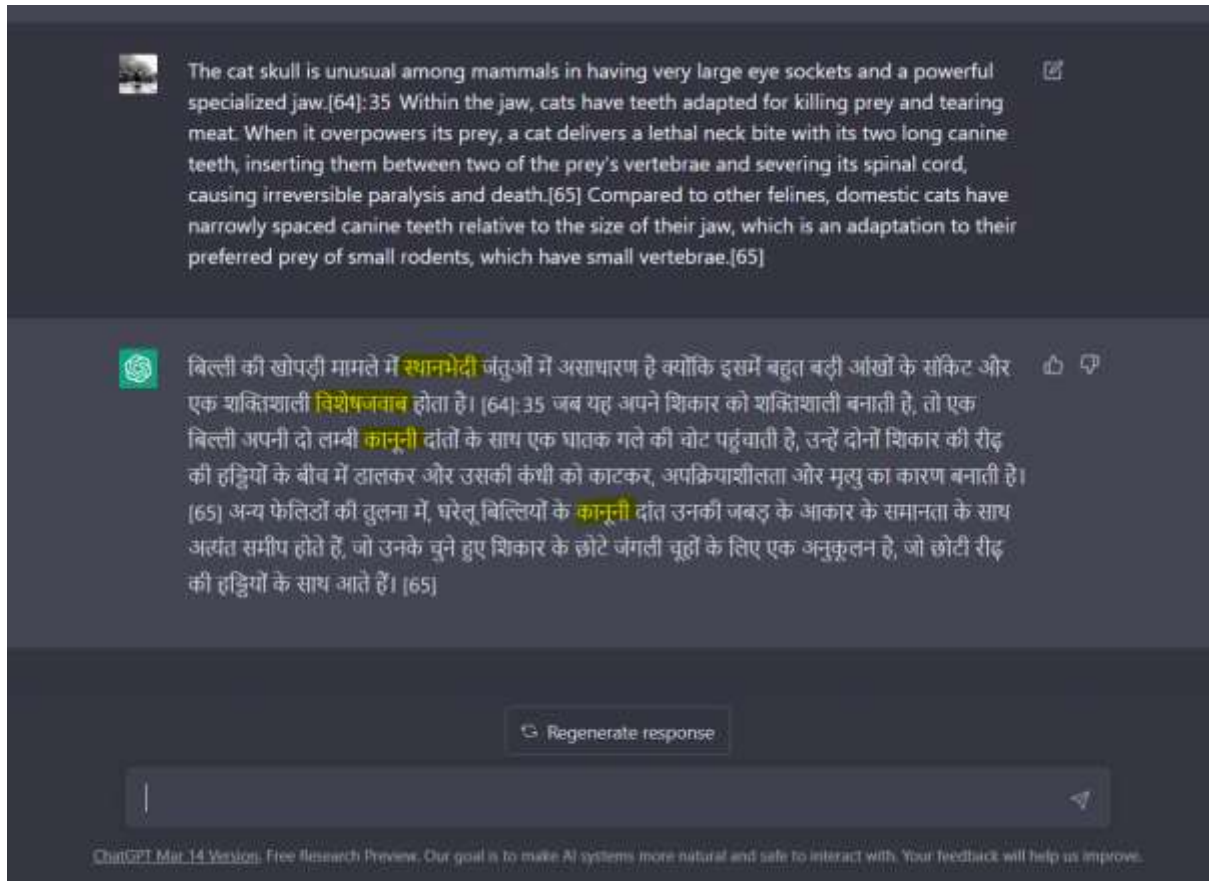
Actual Summary: "Two imprisoned men bond over a number of years, finding solace and eventual redemption through acts of common decency."

Hallucinated Summary: "The Shawshank Redemption is a story about a group of prisoners who escape from jail and go on a wild adventure to find a hidden treasure."

Here in Hallucinated Summary fact like: “go on a wild adventure to find a hidden treasure” has been mentioned which is not supported by the source.

## Task 2 – Machine Translation (English to Indian Language (Hindi) and Vice-versa)

### a. English to Hindi:



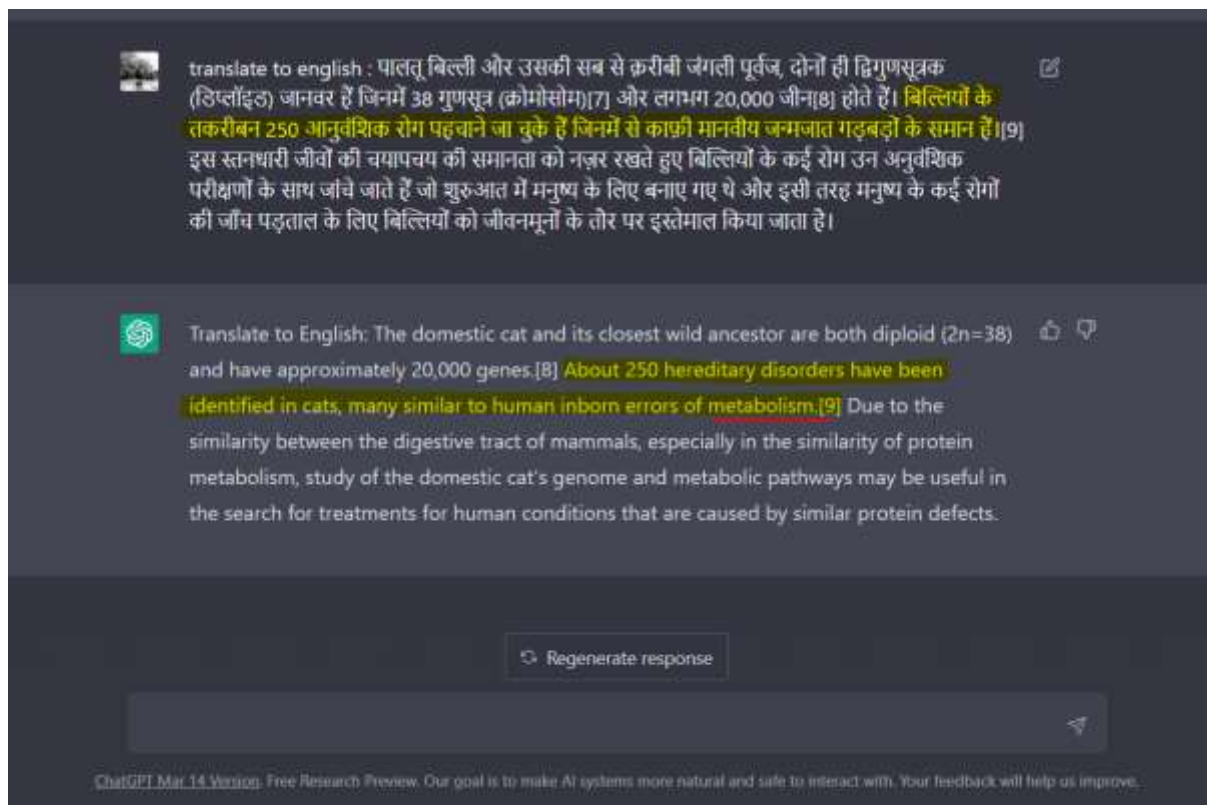
The screenshot shows a ChatGPT interface. The input text in English is: "The cat skull is unusual among mammals in having very large eye sockets and a powerful specialized jaw.[64]:35 Within the jaw, cats have teeth adapted for killing prey and tearing meat. When it overpowers its prey, a cat delivers a lethal neck bite with its two long canine teeth, inserting them between two of the prey's vertebrae and severing its spinal cord, causing irreversible paralysis and death.[65] Compared to other felines, domestic cats have narrowly spaced canine teeth relative to the size of their jaw, which is an adaptation to their preferred prey of small rodents, which have small vertebrae.[65]". The output text in Hindi is: "बिल्ली की खोपड़ी मामले में स्थानभेदी जंतुओं में असाधारण है क्योंकि इसमें बहुत बड़ी आंखों के सॉकेट और एक शक्तिशाली विशेषजवाब होता है। [64]: 35 जब यह अपने शिकार को शक्तिशाली बनाती है, तो एक बिल्ली अपनी दो लंबी कानूनी दांतों के साथ एक घातक गले की चोट पहुंचाती है, उन्हें दोनों शिकार की रीढ़ की हड्डियों के बीच में डालकर और उसकी कंधी को काटकर, अपक्रियाशीलता और मृत्यु का कारण बनाती है। [65] अन्य फेलिडों की तुलना में, घरेलू बिल्लियों के कानूनी दांत उनकी जबड़ के आकार के समानता के साथ अत्यंत समीप होते हैं, जो उनके चुने हुए शिकार के छोटे जंगली चूहों के लिए एक अनुकूलन है, जो छोटी रीढ़ की हड्डियों के साथ आते हैं। [65]". Below the output is a "Regenerate response" button and a text input field. At the bottom, there is a footer: "ChatGPT Mar 14 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve."

### Correct Translation:

स्तनधारियों के बीच बिल्ली की खोपड़ी बहुत बड़ी आंख सॉकेट और एक शक्तिशाली विशेष जबड़ा होने के कारण असामान्य है। [64]:35 जबड़े के भीतर, बिल्लियों के दांत शिकार को मारने और मांस को फाड़ने के लिए अनुकूलित होते हैं। जब यह अपने शिकार पर हावी हो जाता है, तो एक बिल्ली अपने दो लंबे नुकीले दांतों से एक घातक गर्दन काटती है, उन्हें शिकार के दो कशेरुकाओं के बीच में घुसा देती है और उसकी रीढ़ की हड्डी को अलग कर देती है, जिससे अपरिवर्तनीय पक्षाघात और मृत्यु हो जाती है। [65] अन्य बिल्ली के समानों की तुलना में, घरेलू बिल्लियों में उनके जबड़े के आकार के सापेक्ष कैनाइन दांतों के बीच संकरी दूरी होती है, जो छोटे कृन्तकों के उनके पसंदीदा शिकार के लिए एक अनुकूलन है, जिसमें छोटे कशेरुक होते हैं। [65]

Correct translation does not have the words similar to स्थानभेदी, विशेषज्ञवाब, कानूनी in it, which is present in the translation by ChatGPT. These words are neither relevant nor meaningful with the given context.

b. Hindi to English:



Correct English Translation:

Both the domesticated cat and its closest wild ancestor are diploid animals with 38 chromosomes [7] and approximately 20,000 genes [8]. About 250 genetic diseases have been identified in cats, many of which are similar to human congenital disorders. [9] Given the metabolic similarity of this mammal, many diseases in cats are tested with genetic tests that were originally designed for humans, and cats are used as biological samples for many diseases in humans. is used as.

Correct translation does not have the irrelevant word **metabolism** in it, but which is present in translation by ChatGPT.

Q 3. Explain the types of hallucinations observed by you in your answers, along with possible reasons for them. (5 marks)

**Answer.**

These are the types of hallucinations observed in above mentioned cases:

**Task 1 – Abstractive summarization**

Here in Hallucinated Summary has mentioned fact like: “go on a wild adventure to find a hidden treasure”, which is not supported by the source, rather contradicts. So this can be classified as Intrinsic Hallucination.

**Task 2 – Machine Translation (English to Indian Language (Hindi) and Vice-versa}**

a. English to Hindi:

स्थानभेदी, विशेषजवाब, कानूनी words came up in the NLG translated version of the source which are mostly meaning less and irrelevant to the context. Neither contradicting nor supporting the facts mentioned in the source. So, these cases can be classified as instances of **Extrinsic Hallucinations**.

b. Hindi to English:

The word **metabolism** came up in the NLG translated version of the source which is irrelevant to the context. It is neither contradicting nor supporting the facts mentioned in the source. So, it can be classified as instances of **Extrinsic Hallucinations**.

Q 4. Discuss what kind of measures have been proposed in literature about assessing the extent of hallucinations for the tasks of your choice. Also discuss the mitigation methods. (Use pointers from the attached paper.) (5 marks)

**Answer.**

The measures proposed for assessing the extent of hallucinations and mitigation methods for the selected tasks are given below:

## 1. **Abstractive Summarization:**

### 1. a) Measures for assessing the extent of hallucinations

(Page no 17 – point no 7.2)

The paper divides the hallucination metrics for abstractive summarization into two categories: unsupervised and semi-supervised. Existing metrics evaluate intrinsic and extrinsic hallucinations together in one metric because it is challenging to automatically distinguish between them.

#### 7.2.1 Unsupervised Metrics:

- Information extraction (IE)-based metrics
- Natural language inferencing (NLI)-based metrics
- Question answering (QA)-based metrics

IE-based metrics leverage IE models to extract knowledge from both the generation and knowledge source to analyse the factual accuracy of the generation. However, IE models are not reliable, so an entity-based metric relying on the Named-Entity Recognition model is proposed.

NLI-based metrics can be utilized to measure hallucination based on the assumption that a faithful summary will be entailed by the gold source. To improve NLI models for hallucination evaluation, collected annotations are released as additional test data. Other efforts have also been made to further improve NLI models.

QA-based metrics measure the knowledge overlap or consistency between summaries and source documents. QA-based metrics follow three steps to obtain a final score: generating questions from the summaries, obtaining answers from the source documents, and comparing the set of answers from source documents and summaries.

7.2.2 Semi-Supervised Metrics: Semi-supervised metrics are trained on synthetic data generated from summarization datasets. FactCC is a weakly supervised model proposed for evaluating factual consistency. Zhou et al. introduce a method to fine-tune a pre-trained language model on synthetic data with automatically inserted hallucinations to detect the hallucinatory content in summaries.

### 1. b) **Mitigation** methods for addressing hallucinations in abstractive summarization: (Page no 19, Point no 7.3)

Three approaches have been proposed to reduce the hallucination phenomenon in abstractive summarization in this paper.

### **7.3.1 Architecture Method:**

Researchers have made modifications to seq-to-seq models for abstractive summarization to reduce hallucinated content in summaries. They have improved the encoder, decoder, or both, with techniques such as using a graph neural network to encode fact tuples, incorporating external knowledge from Wikipedia, and introducing the Focus Attention Mechanism. Cao et al. and Li et al. have also proposed methods to improve encoder-decoder models with multi-task learning and fact descriptions extraction.

### **7.3.2 Training Method:**

Some researchers have improved the training approach for summarization models to reduce hallucination. Cao and Wang used a contrastive learning method with positive training data being reference summaries and negative training data being hallucinatory summaries. Tang et al. proposed a contrastive fine-tuning strategy called CONFIT for dialogue summarization to improve the factual consistency and overall quality of summaries.

### **7.3.3 Post-Processing Method:**

Several methods have been proposed to reduce the inaccuracies in model-generated summaries, known as draft summaries. One approach is to use post-editing correction models like SpanFact and Cao et al.'s corrector module to correct factual errors in the generated summaries. HERMAN is a system that recognizes and verifies quantities in the summaries for factual consistency. Another method is the contrast candidate generation and selection system proposed by Chen et al., which replaces named entities in the summaries with ones from the source documents and selects the best candidate as the final output summary.

## **2. Machine Translation:**

### **2. a) Measures for assessing the extent of hallucinations**

(Page no 30 – point no 11.2)

Researchers have difficulty defining and detecting hallucinations in machine translation due to its qualitative and subjective nature, but there have been attempts to use statistical methods to automate and quantify the search for hallucinations. It's important to note that the appearance of hallucinations does not seem to affect the BLEU score of the translated text.

#### **11.2.1 Statistical Metrics:**

Martindale et al. propose a metric called bag-of-vectors sentence similarity (BVSS) to identify sentence adequacy. This metric measures the amount of information lost or gained between the reference and machine translation output, indicating if the reference has more information or if the MT output has more information.

### 11.2.2 Model-Based Metrics:

- **Auxiliary Decoder:** Feng et al. propose an "evaluation decoder" to measure faithfulness/adequacy based on word-by-word translation probabilities. The evaluation module adjusts the probability returned by the translation module by returning loss.
- **Entropy Measure:** Tu et al. and Garg et al. use an entropy measure of average attention distribution to detect hallucinations. Hallucinations are visible in attention matrices, and the entropy is calculated on the average attention weights when the model does or does not produce hallucinations during testing.
- **Token Level Hallucination Detection:** Zhou et al. propose a method to detect hallucinated tokens within a sentence. They use a synthetic dataset and compute the hallucination prediction loss between binary labels and the tokens from the hallucinated sentence. They employ similarity-based and overlap-based methods as baselines for hallucination.
- **Similarity-based Methods:** Zhou et al. use an unsupervised model that extracts alignments from similarity matrices of word embeddings and predicts the target token as hallucinated if it is not aligned to the source. Parthasarathi et al. propose computing faithfulness by computing similarity scores between perturbed source sentence and target sentence after applying the same perturbation.
- **Overlap-based Methods:** Zhou et al. predict that the target token is hallucinated if it does not appear in the source. Kong et al. suggest the Coverage Difference Ratio (CDR) as the metric to evaluate adequacy. It is estimated by comparing source words covered by generated translation with human translations.
- **Approximate Natural Hallucination Detection:** Raunak et al. propose an Approximate Natural Hallucination (ANH) detection method based on the fact that hallucinations often occur as oscillations and the lower unique bigram count indicates a higher appearance of oscillatory hallucinations. Their method finds translation above a certain n-gram threshold and searches for repeated targets in the output translation.

## 2. b) **Mitigation** methods for addressing hallucinations in Machine Translation: (Page no 32, Point no 11.3)

The passage discusses the challenges of detecting hallucinations in neural machine translation (NMT) and the importance of mitigating them to improve NMT performance and reduce errors and dangers. NMT engines such as Google and Baidu are widely accessible, leading to a significant interest in improving their performance. The passage outlines methods for mitigating hallucinations in NMT.

### 11.3.1 Data-Related:

1. Data augmentation is a commonly used method, which involves adding perturbed sentences to the dataset. The most successful perturbation method involves inserting the most common tokens at the beginning of the sentence.
2. Corpus filtering is another method to remove hallucinations caused by noise in the dataset. This involves removing repetitive and mismatched source and target sentences.
3. Junczys-Dowmunt proposes a cross-entropy data filtering method for bilingual data that calculates cross-entropy scores for noisy pairs according to two translation models trained on clean data. The scores that suggest disagreement between sentence pairs from the two models are penalized.
4. The influence of fine-grained semantic divergences on NMT outputs is analysed, and a mitigation method based on semantic factors is proposed in [15]. Tags are applied to each source and target sentence to inform about the position of divergent tokens. Factorizing divergence not only helps to mitigate hallucinations but also improves the overall performance of the NMT. Tagging small semantic divergences can provide useful information to the network during training.

### 11.3.2 Modeling and Inference:

The article discusses various methods to mitigate the problem of overexposure bias in Neural Machine Translation (NMT). Wang and Sennrich propose minimum risk training (MRT) as a training objective instead of maximum likelihood estimation (MLE). Scheduled sampling and differentiable approximation to greedy decoding are classic methods used to mitigate this problem. Zhou et al. propose a method to improve self-training of NMT based on hallucination detection, while Li et al. propose tilted empirical risk minimization (TERM) as a training objective. Techniques such as dropout, L2E regularization, and clipping can reduce the number of hallucinations. Additionally, several authors propose methods to improve phrase alignment for better translation accuracy and content identification.



Q 5. Imagine your role as an architect for an NLG system that is being planned to understand the utility of the Large Language Models (LLM) for the organization i.e. planning a “Proof of Concept”. You have the freedom to choose any functionality for the organization and show how LLMs can be used advantageously. You can be highly innovative or highly utilitarian. (5 marks)

- (a). What is the system you are going to propose? Why? Which functionality is it addressing?
- (b). Explain the high-level architecture of your system – marking the inputs and expected outputs of the system.
- (c). What would be your test plan for the system? How would you evaluate it?
- (d). Are there any safeguards that you will suggest to the organization before adopting your system? Justify your answer

**Answer.**

**a.**

The NLG system I would propose is a personalized multilingual customer support system that uses Large Language Models (LLMs) to analyse customer requirement and concern and prepare response for individual customer's preferred language.

This system addresses the functionality of enhancing the customer experience for the organization. For India like countries where so many linguistic groups are present, it is very hard to reach out all the customers in their native language. Multilingual customer support is particularly important in India, which is a diverse country with many different languages spoken. Natural language generation (NLG) technology like ChatGPT can be used to provide multilingual customer support in India. Here's how:

1. Question answering in different languages: ChatGPT can be trained to understand and respond to queries in different languages spoken in India, such as Hindi, Bengali, Tamil, Telugu, and others. The more languages ChatGPT is trained in, the better it can understand and respond to queries in different languages. For instance, when a customer submits a query in a particular language, the chatbot can automatically switch to that language and respond in the same language. Chatbots like ChatGPT can also translate queries and responses into different languages, making it easier for customers to communicate with the support team.
2. Voice-based support: NLG can also be used to provide voice-based support in different languages. ChatGPT can be trained to understand and respond to voice-based queries in different languages, making it easier for customers who are not comfortable with typing.
3. NLG for translations: NLG can be used to generate translations of queries and responses in different languages. This can be particularly useful when the support team needs to communicate with customers who speak a different language.

Overall, NLG technology like ChatGPT can be an effective way to provide multilingual support to customers in India. By using ChatGPT and other NLG tools, customer support teams can provide faster, more efficient, and personalized support to customers in the language they are most comfortable with.

## **b.**

### Inputs:

- Customer queries or issues submitted in different languages spoken in India (e.g. Hindi, Bengali, Tamil, Telugu, etc.)
- User preferences, such as preferred language, communication channels, etc.

### Outputs:

- Responses to customer queries or issues in the language they submitted their query in or their preferred language
- Translation of customer queries or responses into different languages, as needed
- Personalized responses based on user preferences

### High-Level Architecture:

1. User input: The customer submits their query or issue in their preferred language via a chat interface.
2. Chatbot interaction: The chatbot powered by NLG technology like ChatGPT, receives and understands the query using natural language processing (NLP).
3. Language detection: The chatbot detects the language in which the query was submitted using a language detection algorithm.
4. Language-specific processing: The chatbot processes the query in the detected language using a language-specific model trained on ChatGPT.
5. Response generation: The chatbot generates a response in the same language as the query or the customer's preferred language.
6. Translation (if needed): If the response needs to be translated into a different language, ChatGPT can generate a translation of the response.
7. User preference: If the customer has indicated a preference for a particular communication channel or type of response, the chatbot can use this information to provide a personalized response.
8. Delivery: The response is delivered to the customer via the chat interface.

This high-level architecture demonstrates how multilingual customer support can be provided using NLG technology like ChatGPT. The inputs to the system include customer queries or issues submitted in different languages, and the expected outputs are responses in the same language or a preferred language, translations, and personalized responses. The chatbot powered by ChatGPT processes the input, generates a response, and delivers it to the customer via a chat interface.

## C.

Here's a high-level test plan for evaluating the multilingual customer support system using NLG technology like ChatGPT:

1. Functional testing: Test the system's functionality to ensure that it can understand and respond to queries in different languages, provide personalized responses based on user preferences, and translate responses when needed.
2. Performance testing: Test the system's performance to ensure that it can handle multiple concurrent queries, respond quickly, and maintain uptime during peak traffic.
3. User acceptance testing: Test the system with a group of end-users who speak different languages to ensure that it meets their needs, is easy to use, and provides a satisfactory user experience.
4. Security testing: Test the system's security features to ensure that it can handle sensitive customer information, and that it is resistant to attacks such as SQL injection, cross-site scripting, and others.
5. Load testing: Test the system's ability to handle different volumes of traffic and queries without crashing, slowing down, or experiencing other performance issues.
6. Integration testing: Test the system's ability to integrate with other systems such as customer relationship management (CRM) software, payment gateways, and other third-party applications.
7. Localization testing: Test the system's ability to adapt to different cultures and languages by testing the translations, cultural appropriateness, and sensitivity of responses.
8. Accessibility testing: Test the system's accessibility features to ensure that it can be used by people with different abilities, including those with visual or hearing impairments.

To evaluate the system, we can use both manual and automated testing methods. Manual testing can be done by trained testers who can interact with the system in different languages and scenarios, and evaluate its performance, functionality, and usability.

However, evaluation of NLG model can be done in following ways:

1. **Human evaluation:** This technique involves having human evaluators rate the quality of the system's generated answers. Evaluators can be asked to rate the accuracy, relevance, fluency, and overall quality of the answers on a scale of 1-5. Human evaluation can provide valuable insights into the strengths and weaknesses of the system's answers, and can also be used to fine-tune the system's performance.
2. **Automatic evaluation:** This technique involves using automated metrics to evaluate the quality of the system's generated answers. Commonly used metrics include BLEU, ROUGE, METEOR, and others. These metrics compare the system's generated answers to a reference set of answers and assign a score based on their similarity. While automatic evaluation can provide quick and objective results, it may not always capture the nuances and context of the generated answers.
3. **Domain-specific evaluation:** This technique involves evaluating the system's performance on a specific domain or task, such as answering questions related to a specific topic or industry. This can be done by testing the system's accuracy, relevance, and fluency in answering questions within that domain. Domain-specific evaluation can provide insights into the system's ability to handle different types of questions and topics.
4. **Error analysis:** This technique involves analysing the errors made by the system in generating answers. This can help identify common errors, patterns, and areas for improvement. Error analysis can be done manually by analysing a sample of generated answers or using automated tools that can identify common errors.
5. **User feedback:** This technique involves soliciting feedback from users who have interacted with the system. This can be done through surveys, interviews, or other forms of feedback mechanisms. User feedback can provide insights into the user's satisfaction with the system's generated answers, ease of use, and overall experience.

By using a combination of these techniques, we can evaluate the performance of a Generative Question Answering system powered by NLG technology. This evaluation can help identify areas for improvement and fine-tune the system's performance to better meet the user's needs.

**d.**

Adopting multilingual customer support using NLG technology like ChatGPT in India can bring many benefits to an organization, but it is important to consider the potential risks and put safeguards in place. Here are some safeguards that can be suggested:

1. **Data privacy and security:** Ensure that the system is secure and meets data privacy regulations. This includes using encryption, access controls, and data backup procedures to protect customer data.
2. **Accuracy and fairness:** Verify the accuracy and fairness of the system's responses, particularly for sensitive topics. It is important to ensure that the system does not generate biased or discriminatory responses.
3. **Transparency:** Ensure that the system's operations and decision-making processes are transparent to users. This includes providing information about how the system generates responses, what data is used, and how it handles user data.
4. **User consent:** Obtain user consent for collecting and using their data, and provide them with the option to opt-out or delete their data if they wish.
5. **Human oversight:** Provide human oversight to the system's operations, particularly for complex or sensitive queries. This can include having a team of human operators who can review the system's responses and provide feedback.
6. **Continuous monitoring and evaluation:** Continuously monitor and evaluate the system's performance, particularly for accuracy and fairness. This can include regular testing, user feedback, and error analysis.
7. **Technical support:** Provide technical support to users who may have difficulty using the system, particularly for users who are not proficient in the languages used by the system.

By implementing these safeguards, organizations can minimize the potential risks associated with adopting multilingual customer support using NLG technology like ChatGPT in India. This can help ensure that the system is secure, accurate, fair, and transparent, and that it provides a satisfactory user experience.