ATLANTIS at SemEval-2025 Task 3: Detecting Hallucinated Text Spans in Question Answering

Catherine Kobus, Francois Lancelot,
Marion-Cecile Martin, Nawal Ould Amer
Airbus AI Research

Abstract

This paper presents the contributions of the ATLANTIS team to SemEval-2025 Task 3, focusing on detecting hallucinated text spans in question answering systems. Large Language Models (LLMs) have significantly advanced Natural Language Generation (NLG) but remain susceptible to hallucinations, generating incorrect or misleading content. To address this, we explored methods both with and without external context, utilizing few-shot prompting with a LLM, token-level classification or LLM fine-tuned on synthetic data. Notably, our approaches achieved top rankings in Spanish and competitive placements in English and German. This work highlights the importance of integrating relevant context to mitigate hallucinations and demonstrate the potential of fine-tuned models and prompt engineering.

1 Introduction

LLMs have achieved remarkable proficiency in NLG, enabling significant improvements across various applications, including translation (Alves et al., 2024), classification (Li et al., 2023), synthetic data generation (Dai et al., 2022), Retrieval-Augmented Generation (RAG) systems (Seo et al., 2024). While they have addressed numerous challenges in these domains, they remain prone to hallucination-generating incorrect or misleading content. This issue can undermine system reliability and negatively affect real-world performance, limiting their practical deployment in critical applications.

To tackle this challenge, the SemEval MuSHROOM task focuses on detecting hallucinated spans in generated text, a crucial step toward enhancing the trustworthiness of NLG systems. This multilingual task covers 14 languages and requires identifying specific portions of text where hallucinations occur. The task overview paper (Vázquez et al., 2025) provides a comprehensive analysis of the methodologies and findings, offering valuable insights into hallucination detection in NLG systems. The dataset, the evaluation metrics, and models used in the task are also extensively discussed there.

In this paper, we present a set of approaches for the challenge, which fall into two main categories:
Methods without external context that solely rely on the question and answer as input. We experimented 1) few-shot prompting using LLM and 2) fine-tuning a token-level classifier on our generated synthetic data using MKQA dataset (Longpre et al., 2020).
Methods with external context from Wikipedia, retrieved using our RAG system. This context is then added into our models in three ways: 1) few-shot prompting with LLM, 2) fine-tuning a token-level classifier, and 3) fine-tuning a LLM for hallucinated span detection.

Our models delivered impressive results in several languages. In particular, we ranked first in Spanish, third in English, and fifth in German using few-shot prompting with Gemini Pro, enhanced by contextual information. For French, we achieved the eleventh place with a fine-tuned token classifier model with context.

2 Systems overview

In this section, we outline our different approaches. We begin by introducing the retrieval component of our RAG system. Next, we describe our methods based on few-shot prompting with and without retrieval. Finally, we present the approaches that we fine-tuned for the task using synthetic data.

2.1 Retrieval module

The retrieval module is designed to extract relevant text segments to answer a given question. We used the Wikipedia dataset[1] from November 2023 as

[1] https://hf.co/datasets/wikimedia/wikipedia

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1098–1107
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

source. The text was chunked into segments of 312 tokens each, with an overlap of 100 tokens, resulting in 21 million indexed chunks. These chunks were indexed using both dense representation with BAAI/bge-large-en-v1.5 embedding model[2] and sparse representation with BM25.

The retrieval process comprises three key steps: retrieval, reranking, and clustering. During the retrieval step, a hybrid search mechanism selects the top 25 chunks. This hybrid search employs both an embedding model and BM25 with distribution-based score fusion (Mazzeschi, 2023). Then, a cross-encoder model[3] reranks these 25 chunks. Based on the computed reranker scores, a k-means clustering algorithm is applied to retain a variable number of the most relevant chunks.

To support multiple languages of the query, we employed a LLM (Mistral-7B-Instruct-v0.2[4]) to translate the query into English. This translation allows the retrieval of relevant Wikipedia context in English from a question in another language.

2.2 Approaches without additional fine-tuning

We evaluated two approaches that do not require additional fine-tuning: an overlap-based baseline method and a LLM with a custom prompt.

2.2.1 Overlap-based method

The overlap-based method is a heuristic approach that predicts a target token as hallucinated if it does not appear in the context; it is inspired from the overlap-based method detailed in (Zhou et al., 2020). Since only the English version of Wikipedia was indexed, this method was tested exclusively for the English language.

2.2.2 LLM and prompt engineering

Prompt Engineering. A more flexible approach leverages LLMs, which have demonstrated strong adaptability across various tasks through prompt engineering. To address our challenge, we designed a custom prompt tailored specifically for this task. The LLM was provided with two examples and instructed to generate responses in a structured format—JSON in our case, as illustrated in the prompt 1 in the appendix. Our objective was to maximize the capabilities of an LLM by first detecting hallucinations, then implement custom functions to extract the hallucinated spans and identify their positions within the sentence.

With retrieval. After testing fixed prompt strategies, we incorporated textual evidence in the prompt. Providing relevant documents has been shown to reduce hallucination. Moreover, in the challenge setup, it allows for a direct comparison between facts and the answer to be evaluated. The relevant chunks are extracted from Wikipedia English and selected for each question as described in 2.1.

To balance between LLM prior knowledge and additional knowledge, rules with different degrees of strictness have been explored, inspired from (Wu et al., 2024). Keeping the flexibility to rely on prior knowledge was important for cases where the retrieval pipeline was unable to find documents with relevant facts or when conflictual information was present in the given chunks. Stricter rules also helped to highlight the minimal hallucinated part in the answer. Finally, we tested with including misspellings, such as "Stoveren" instead of "Staveren" for the first example of the English validation set. However, this type of errors was often not labeled in the challenge dataset, therefore we discarded them.

For the first stage of our experiments, we tested on the English dataset only. To adapt to German, French and Spanish, we simply named the language in the prompt and changed one example with question and answer in this other language. Listing 2 shows the prompt used in the multilingual setting.

Experimental setup. The Gemini 1.5 Pro model (Team et al., 2024) was prompted with a fixed seed and a temperature of 0.0 to foster the replicability of the results. Given the large context size, all the chunks tagged as relevant could be incorporated in the prompt: it represents between 1 and 23 chunks per question, with a median from 4 chunks for French to 6 for Spanish.

2.3 Approaches with additional fine-tuning

The challenge dataset did not provide annotated training data - only small annotated validation dataset. Therefore, we created synthetic data to fine-tune custom models. Using this, we fine-tuned a token-level classifier described in 2.3.2 and a LLM described in 2.3.3.

2.3.1 Data generation process

LLMs have become increasingly popular for synthetic data generation in various NLP applications

[2] https://hf.co/BAAI/bge-large-en-v1.5

[3] https://hf.co/BAAI/bge-reranker-large

[4] https://hf.co/mistralai/Mistral-7B-Instruct-v0.2

(Liu et al., 2024; Seo et al., 2024). To generate our data, we used MKQA dataset (Longpre et al., 2020), and excluding long-answer or unanswerable queries. Given a question and context retrieved using the retrieval module described in 2.1, we prompted a LLM (Gemini 1.5 Pro) to generate a short answer and an answer repeating the question to have a format closer to the challenge dataset. Table 4 shows an example. Once we get the answer, we prompt the LLM to inject a hallucination, with few-shot learning that provides guidance through examples. The resulting generated dataset contains around 48000 samples (12000 samples per language). Appendix 5 shows two generated samples.

2.3.2 Token-level classification

The span hallucination task can be also casted as a more classical token classification task, where each token in the LLM output is assigned a label, either '**I-H**' (if the token is part of an hallucination) or '**O**' (outside an hallucination). This approach takes inspiration from the XLM-R baseline provided by the challenge organizer and from the hallucination detection method for Machine Translation described in (Zhou et al., 2020). The architecture of the approach is illustrated in Figure 1 with an example taken from the English validation set provided in the challenge. A linear layer is added on top of the pretrained XLM-RoBERTa[5] model in order to perform the classification at token level.

We used the synthetic data generated following the procedure detailed in 2.3.1 as training data. Different configurations were tested in the course of the challenge, with or without providing the relevant Wikipedia chunks from 2.1, putting the question either before, after the relevant context or omitting it. For this last configuration, experiments showed that putting the question at the beginning leads to better performances.

Since XLM-RoBERTa has a maximum sequence length of 512 tokens, we only provide the top-1 retrieved chunk of document as input context to the model. By doing so, the total input length to the model (including, at most, the question, a Wikipedia chunk, and the LLM output) never exceeded the model's maximum sequence length.

During the challenge, different fine-tunings in monolingual and multilingual mode were explored; more details can be found in A.1.

Experimental setup. The XLM-RoBERTa large model was fine-tuned for 7 epochs on 4 A10G GPUs, with a batch size of 6 and a learning rate of 2×10^{-5}. In the multilingual setting, the training/development data sizes are respectively around 44000/4000 examples, while in the monolingual setting, the training/development data sizes are respectively around 11000/1000 examples.

The checkpoint that gave the best result on the challenge validation set was selected for the test. We also tuned the probability threshold for the '**I-H**' class. By default, the decision threshold is 0.5; however, we noticed that the model was globally under-confident, and, by decreasing the threshold, we could increase the results in terms of IoU.

2.3.3 Fine-tuned LLM

We fine-tuned a LLM for hallucination detection, based on the work of (Mishra et al., 2024) which introduced FAVA (FAct Vericaton with Augmentation), a model for fine-grained hallucinations detections and editing. We adapted this approach to our question-answering task: we modified the training data to include the question along with the context and answer. Additionally, we simplified the training data by focusing on a single type of hallucinated entity (instead of the six presented in the original paper). We also fine-tuned the LLM with the multilingual synthetic data detailed in 2.3.1 compared to the initial FAVA model which was only fine-tuned for English.

Listing 4 shows one sample used for the fine-tuning of the LLM with a French question, the retrieved context in English, and the edited output to correct the hallucination.

Experimental setup. We fine-tuned a Llama-3.2-3B-Instruct model for 2 epochs with LoRA, using a batch size of 36, with $rank = 128$ and $\alpha = 128$, 4-bit quantization, the Adam optimizer and a learning rate of 2×10^{-4} on 1 A10G GPU.

3 Results

3.1 Performance of the retrieval system

Table 1 summarizes the retrieval results obtained on the test set. As only the English articles were indexed, we computed the retrieval performance for the other languages by converting the "English" retrieved Wikipedia URL into the target language URL using the MediaWiki API[6] that links equivalent Wikipedia pages in different languages.

[5] https://huggingface.co/FacebookAI/xlm-roberta-large

[6] https://www.mediawiki.org/wiki/API:Langlinks

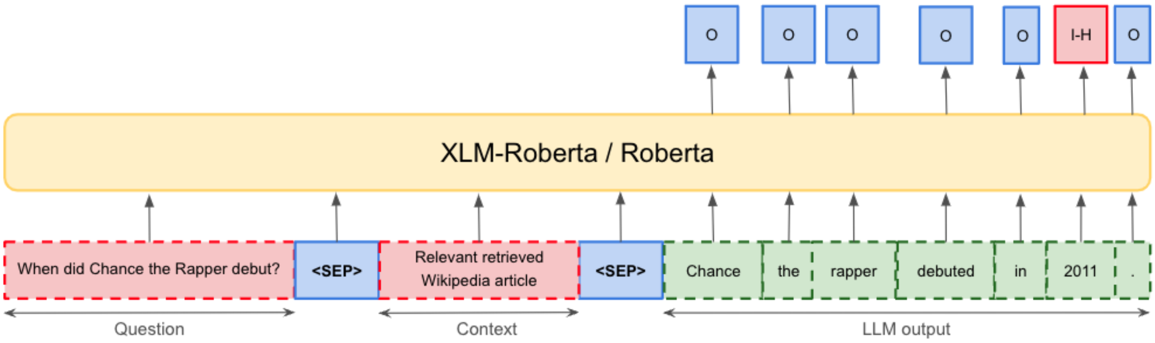

Figure 1: Architecture of the token classification approach illustrated with an example

language	en	de	es	fr
test	0.81	0.63	0.70	0.62

Table 1: Retrieval scores (MAP@5).

The model performs best in English, with scores (MAP@5) around 0.80. The lower retrieval performance in other languages (German, Spanish, and French) can be attributed to the additional translation step and the lack of corresponding English-indexed Wikipedia articles for some relevant articles in those languages.

3.2 Performance of the hallucination detection

Table 2 summarizes the results obtained on the test set for four languages: English, German, Spanish and French.

3.2.1 Performance comparison without retrieval

In this section, we compare the performance of our strategies in a setting without retrieval, thereby evaluating their standalone capabilities for hallucination detection without external knowledge augmentation.

As reported in Table 2, the XLM-RoBERTa large model exhibits moderate performance across languages, achieving its highest test IoU score of 0.50 in French, and its lowest in Spanish, 0.27. In contrast, Gemini 1.5 Pro demonstrates competitive overall performance, outperforming XLM-RoBERTa in German (0.45 vs. 0.38) but underperforming in French (0.43 vs. 0.50).

These findings suggest that, despite the larger scale of general-purpose models like Gemini, smaller models that have been fine-tuned on task-specific data can yield comparable results. Moreover, given that Gemini is pre-trained on general data, our hypothesis is that the fine-tuned XLM-RoBERTa large model would likely exhibit superior performance in domain-specific applications.

3.2.2 Performance comparison with retrieval

Here, we focus on the performance of our strategies in a setting with retrieval, thereby evaluating their capabilities for hallucination detection using external knowledge (RAG). Comparing with the previous section, the scores are better in all languages, without exception.
Gemini 1.5 Pro still outperforms the fine-tuned approaches on 3 over 4 languages in this setting. However, the finetuned Llama-3.2-3B reaches the same average IoU of 0.54 accross all languages, notably given the size difference of these models. Morevover, XLM-RoBERTa, significantly smaller, achieves a score that is relatively close to Llama-3.2-3B in German (0.53 vs 0.57).
These findings suggest that fine-tuning on synthetic data is a promising strategy and leveraging robust retrieval mechanisms with diverse pre-training can yield superior performance in the complex task of hallucinated span extraction.

3.3 Limitations

Our approaches have several limitations.

For the retrieval-based method, we indexed only the English version of Wikipedia. If relevant facts reside in other sources, the information retrieval (IR) system cannot provide the necessary context. Additionally, we relied on a LLM to translate queries into English before retrieval. This approach could have been compared with multilingual embedding models and vocabulary-based retrieval to evaluate its effectiveness.

Our prompt-based methods required extensive manual experimentation to design prompts that

	en	de	es	fr	Avg.
Without retrieval					
Finetuned XLM-RoBERTa large*	0.42	0.38	0.27	0.50	0.39
Gemini 1.5 Pro	0.40	0.45	0.34	0.43	0.41
With retrieval					
Overlap-based	0.36	-	-	-	-
Finetuned XLM-RoBERTa large*	0.51	0.53	0.37	0.55	0.49
Finetuned Llama-3.2-3B*	0.55	0.57	0.39	**0.63**	0.54
Gemini 1.5 Pro	**0.57**	**0.58**	**0.53**	0.50	**0.54**

Table 2: IoU scores on test set. We **bold** the best performance across submitted systems. Approaches with * were finetuned on a synthetic dataset.

aligned with the characteristics of this challenge's dataset. This process was time-consuming and constrained by the limited number of prompts we could test manually. Finding the most effective prompt remains inherently difficult, and a more systematic approach—such as training a model to optimize prompt selection—could have improved our results.

For the token-level classification model, the limited context window constrained our ability to incorporate all relevant information. Only the first chunk of text was appended to the context, which could be problematic when key details were spread across multiple chunks. A potential solution would be to filter and include only the most relevant sentences to enhance classification accuracy.

Finally, both the token-level classifier and the fine-tuned LLM were trained on synthetic data. Ensuring the accuracy and fidelity of this data is a major challenge. If the synthetic data contains errors, hallucinations, or biases, the trained models may fail to generalize effectively to real-world scenarios, leading to unreliable predictions and reduced robustness (van Breugel et al., 2023). Moreover, the quality of synthetic data depends heavily on the data generation process itself. Addressing these issues would require more rigorous validation techniques or alternative data augmentation strategies to improve the reliability of the training data.

4 Conclusion

Different approaches were presented with and without retrieval for the hallucinated span detection task. Overall, the task remains difficult and the performance of the same strategy varies widely depending on the language. This work underlines the importance of adding a relevant context to detect hallucinated spans in the answer. In general, LLM prompting leads to better results and is easily adaptable on other languages but the smaller fine-tuned models show promising results and could thus be preferred, subject to further tuning. Lastly, these approaches would need to be validated on a balanced dataset, containing also a significant part of non-hallucinated answers.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering.

Michelangiolo Mazzeschi. 2023. Distribution-based score fusion (dbsf), a new approach to vector search ranking.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucinations detections. *arXiv preprint*.

Minju Seo, Jinheon Baek, James Thorne, and Sung Ju Hwang. 2024. Retrieval-augmented data augmentation for low-resource domain tasks.

Gemini Team, Petko Georgiev, and All. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Boris van Breugel, Zhaozhi Qian, and Mihaela van der Schaar. 2023. Synthetic data, real errors: how (not) to publish and use synthetic data.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona de Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 Task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes.

Kevin Wu, Eric Wu, and James Y Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. In *Advances in Neural Information Processing Systems*, volume 37, pages 33402–33422. Curran Associates, Inc.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona T. Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. 2020. Detecting hallucinated content in conditional neural sequence generation. *CoRR*, abs/2011.02593.

A Appendix

A.1 Monolingual/multilingual token-level classification

This sections contains further details about experiments conducted for the token-level classification strategy, especially with respect to fine-tuning with one or more languages.

Experimental setup. At the beginning of the challenge, we focused on English. We finetuned both RoBERTa[7] and XLM-RoBERTa large models on the English synthetic dataset, with the configurations mentioned in 2.3.2. The maximum sequence length for RoBERTa is also 512 tokens, therefore we added the top 1 chunk retrieved when experimenting with retrieval. Then, we decided to extend to other languages of the challenge for which we could create synthetic data - French, Spanish and German. We fine-tuned XLM-RoBERTa in two ways:

- Multilingual: on the aggregated synthetic data for all languages
- Monolingual: for each of the 3 new languages, on the subpart of the synthetic data with the target language

For each language, the best checkpoint was selected and the probability threshold was adapted. Table 3 shows the IoU scores obtained on the challenge test sets for English, German, Spanish and French. For each configuration, the first score is without adding context, and the second one is with additional context.

Results. First of all, the results of 3.2 are validated: adding a relevant context always leads to better performances regardless of the finetuning setting. Monolingual fine-tuning gives higher performance for German (from 0.53 to 0.55) and English (from 0.51 to 0.54) , whereas better results are reached with multilingual finetuning for French (from 0.46 to 0.50) and Spanish (from 0.35 to 0.37), with retrieval. In this case, it seems to benefit from the training with data from other languages.

The preference to fine-tune specifically on a language or on all varies with respect to the language considered, as well as the performance achieved which is significantly lower for Spanish. Further work could focus on optimizing the fine-tuning to reach a single model that performs well across all these languages. For example, one could use knowledge distillation from the best checkpoints by language into a unique model to obtain multilingual capabilities.

A.2 Sample generated data

A.3 Prompts

[7] https://huggingface.co/FacebookAI/roberta-large

Finetuning setting	**Model**	**en**	**de**	**es**	**fr**
Monolingual	Roberta large	**0.45 / 0.54**	-	-	-
	XLM-Roberta large	0.45 / 0.52	**0.44 / 0.55**	**0.29** / 0.35	0.46 / 0.51
Multilingual	XLM-Roberta large	0.42 / 0.51	0.38 / 0.53	0.27 / **0.37**	**0.50 / 0.55**

Table 3: IoU scores without / with retrieval for the token-level classification strategy on the test set. We **bold** the best performance across finetuning setups.

question	when did the first episode of the flash come out
short rag answer	October 7, 2014
rag answer with question	The first episode of The Flash (2014) premiered on October 7, 2014.

Table 4: A sample of synthetic generated answer

question	when did the first episode of the flash come out
short rag answer with hallucination annotations	<entity><mark>October 7, 2014<mark><delete>October 7, 2015<delete><entity>
short rag answer with hallucination	October 7, 2015
mushroom hallucination hard labels	[[0, 15]]
rag answer with question with hallucination annotations	The first episode of The Flash (2014) premiered on <entity><mark>October 7, 2014<mark><delete>October 7, 2015<delete><entity>.
rag answer with question with hallucination	The first episode of The Flash (2014) premiered on October 7, 2015
mushroom hallucination hard labels	[[51, 66]]

Table 5: A sample of synthetic generated hallucination annotation

```
You are an expert evaluator for language models, tasked with identifying hallucinations or errors in their responses. A
hallucination is defined as:
– Incorrect factual information: Content in the model's response that is unrelated, irrelevant, or factually incorrect with respect
 to the documents, based on the question.
– Spelling errors: Any word in the model's response that is misspelled or contains typographical mistakes, including incorrect
names, places, or other terms.

Your task is to analyze the **Model Output Text** written in {LANGUAGE} and classify hallucinations or errors into two
categories:
1. **Factual inaccuracies**
2. **Misspellings**

Ensure the output is in JSON format, following this structure:
'''
{
    "model_input": "<Insert question here>",
    "model_output_text": "<Insert model's response here>",
    "hallucinations": {
        "factual_inaccuracies": [
            "<text_span_1>", "<text_span_2>", ...
        ],
        "misspellings": [
            "<text_span_3>",
            "<text_span_4>",
            ...
        ]
    }
}
'''

### Example 1:
**Model Input**: "Quelle est la capitale de la France ?"
**Model Output Text**: "La capitale de la Grance est Berlin."
**Expected JSON Output**:
'''
{
    "model_input": "Quelle est la capitale de la France ?",
    "model_output_text": "La capitale de la Grance est Berlin.",
    "hallucinations": {
        "factual_inaccuracies": [
            "Berlin"
        ],
        "misspellings": [
            "Grance"
        ]
    }
}
### Example 2:
...

### Task:
Now, evaluate the **Model Output Text**. Identify hallucinations or errors, and classify them as either factual inaccuracies or
 misspellings. The **Relevant Documents** is a list that can contain different documents in English, all independant. If one
doesn't seem relevant, don't take it into account to identify hallucinations.

**Model Input**: {QUESTION}
**Model Output Text**: {MODEL_OUTPUT}

### Remember instruction:
You MUST select only the relevant subparts of the answer (where the error occurs). You MUST split them into the MINIMAL
 possible parts. You MUST exclude stop words.
```

Listing 1: Prompt for generic LLM & multilingual

```
You are an expert evaluator for language models, tasked with identifying hallucinations or errors in their responses. A
hallucination is defined as:
– Incorrect factual information: Content in the model's response that is unrelated, irrelevant, or factually incorrect with respect
 to the documents, based on the question.
– Spelling errors: Any word in the model's response that is misspelled or contains typographical mistakes, including incorrect
names, places, or other terms.

Your task is to analyze the **Model Output Text** written in {LANGUAGE} and classify hallucinations or errors into two
categories:
1. **Factual inaccuracies**
2. **Misspellings**

Ensure the output is in JSON format, following this structure:
'''
{
    "model_input": "<Insert question here>",
    "model_output_text": "<Insert model's response here>",
    "hallucinations": {
        "factual_inaccuracies": [
            "<text_span_1>", "<text_span_2>", ...
        ],
        "misspellings": [
            "<text_span_3>",
            "<text_span_4>",
            ...
        ]
    }
}
'''

### Example 1:
**Model Input**: "Quelle est la capitale de la France ?"
**Model Output Text**: "La capitale de la Grance est Berlin."
**Relevant Documents**: ["The following outline is provided as an overview of and topical guide to Paris: Paris  capital and
most populous city of France, with an area of and an official estimated population of 2,140,526 residents as of 1 January 2019.
 Since the 17th century, Paris has been one of Europe's major centres of finance, commerce, fashion, science, and the arts...."]
**Expected JSON Output**:
'''
{
    "model_input": "Quelle est la capitale de la France ?",
    "model_output_text": "La capitale de la Grance est Berlin.",
    "hallucinations": {
        "factual_inaccuracies": [
            "Berlin"
        ],
        "misspellings": [
            "Grance"
        ]
    }
}
### Example 2:
...

### Task:
Now, evaluate the **Model Output Text**. Identify hallucinations or errors, and classify them as either factual inaccuracies or
 misspellings. The **Relevant Documents** is a list that can contain different documents in English, all independant. If one
doesn't seem relevant, don't take it into account to identify hallucinations.

**Model Input**: {QUESTION}
**Model Output Text**: {MODEL_OUTPUT}
**Relevant Documents**: {CONTEXT}

### Remember instruction:
You MUST select only the relevant subparts of the answer (where the error occurs). You MUST split them into the MINIMAL
 possible parts. You MUST exclude stop words.
```

Listing 2: Prompt for generic LLM - with retrieval & multilingual

```
Given the following question:
{QUESTION}
and the following references:
{CONTEXT}

Identify all the errors in the following passage using the references and the question provided and suggest edits:
{MODEL_OUTPUT}
```

Listing 3: Prompt for finetuned LLM

```
USER : Given the following question:
quand le livre de la jungle de Disney est−il sorti
and the following references:
# The Jungle Book (2016 film)
This film required extensive use of computer−generated imagery to portray the other animals and settings. The Jungle Book
was released in the United States in Disney Digital 3D, RealD 3D, IMAX 3D, D−Box, and premium large formats, on April
15, 2016. It grossed $966 million worldwide, making it the fifth−highest−grossing film of 2016. Amongst Disney's live−
action remakes, this film is the eighth−most−expensive and fifth−highest−grossing readaptation to date. This film won
accolades for achievements in visual effects at the 89th Academy Awards, 22nd Critics' Choice Awards, and 7 ...
Identify all the errors in the following passage using the references and the question provided and suggest edits:
Le film est sorti le 18 octobre 1967.

ASSISTANT : Le film est sorti le <entity><mark>15 avril 2016</mark><delete>18 octobre 1967</delete></entity>.
```

Listing 4: Example of an input/output pair used for the LLM finetuning

```
Translate the following query into english : {QUERY}. Return only the translated query.
```

Listing 5: Prompt used for translation

```
You must follow those examples to generate a new answer with annotations.

Example 1 :
− Question :
What did Petra van Staveren win a gold medal for?
− Answer :
Petra van Stoveren won a gold medal in the 1984 Summer Olympics in Los Angeles, USA.
− Answer with annotations :
Petra van Stoveren won a <entity><delete>gold</delete><mark>silver</mark></entity> medal in the <entity><delete
>1984</delete><mark>2008</mark></entity> Summer Olympics in <entity><delete>Los Angeles, USA</delete><mark>
Beijing, China</mark></entity>.

Example 2 :
− Question :
Which network released the TV series of the The Punisher?
− Answer :
The Punisher network that released this TV show is Netflix.
− Answer with annotations :
The <entity><delete>Punisher</delete><mark>Puncher</mark></entity> network that released this TV show is Netflix.
...
Example :
− Question :
{QUESTION}
− Answer :
{ANSWER}
− Answer with annotations :
```

Listing 6: Prompt used to generate synthetic hallucination