# GIL-IIMAS UNAM at SemEval-2025 Task 3: MeSSI: A Multimodule System to Detect Hallucinated Segments in Trivia-like Inquiries.

**Francisco López-Ponce,[1,2], Karla Salas-Jimenez[1,2],**
**Adrián Juárez-Pérez[1,4], Diego Hernández-Bustamante[3],**
**Gemma Bel-Enguix[1], Helena Gómez-Adorno[3]**
[1] Grupo de Ingeniería Lingüística - UNAM
[2] Posgrado en Ciencias e Ingeniería de la Computación - UNAM
[3] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas - UNAM
[4] Facultad de Ciencias - UNAM
{karla_dsj,francisco.lopez.ponce,danyjuarez99}@ciencias.unam.mx gbele@iingen.unam.mx
helena.gomez@iimas.unam.mx, diegohernandez969@aragon.unam.mx

## Abstract

We present MeSSI, a multi-module system applied to SemEval 2025's task 3: Mu-SHROOM. Our system tags questions in order to obtain semantic relevant terms that are used as information retrieval characteristics. Said characteristics serve as extraction terms for Wikipedia pages that are in turn processed to generate gold standard texts used in a hallucination evaluation system. A part-of-speech-tag based entity comparison was implemented to contrast the test dataset sentences with the corresponding generated gold standards, which in turn was the main criterion to tag hallucinations, partitioned in *soft labels* and *hard labels*. This method was tested in Spanish and English, finishing 18th and 19th respectively on the IoU based ranking.

## 1 Introduction

Given the increasing use of LLMs, the creation of hallucination detection and fact checking systems is of great importance. The Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes (Mu-SHROOM (Vázquez et al., 2025)) focuses on analyzing and tagging LLMs' responses in order to determine which spans of text correspond to hallucinations and incorrect information, based on a system generated probability. Mu-SHROOM's evaluation is carried out in 14 different languages, and performance is measured in two character-level metrics: intersection over union (IoU) of tagged spans (predictions vs gold-labels), and a comparison between assigned probabilities of each span compared to the gold label probability ($\rho$), the main evaluation being IoU.

This article describes a three module pipeline for solving this task: the **M**ultilmodule **S**ystem to Detect Hallucinated **S**egments in Trivia-like **I**nquiries (MeSSI). Our pipeline uses Wikipedia as an information retrieval database. Based on the retrieved texts, a gold standard answer is generated using automatic summarizing systems or LLMs, finally this processed gold standard is compared with each LLM output from the Mu-SHROOM dataset. Given the nature of the questions, a PoST (Part-of-Speech-Tag) based comparison was used in order to generate the final tags of each span. The MeSSI system was tested in English and Spanish, ranking in the top half of all models in English, and just below the half in Spanish (19th out of 44 participants in English, and 18th out of 35 in Spanish).

Working in this task gave us insight regarding the difficulties of creating automated fact checking systems, particularly given the wide variety of questions that these systems have to verify and the vast amount of information needed to cover all possible subjects. Similarly, the span tagging portions of the task also proved to be a challenge since the tagging has to be based on factual differences that don't always translate to syntactic structure or PoST based differences. Our system proved to be language-resource dependent, almost doubling performance in English compared to Spanish, yet managed to surpass each language's baseline. The code can be found in GitHub[1].

## 2 Background

LLM Hallucination and Fact-Checking has become a widely studied area of research with various approaches. Jiang et al. (2020) introduced HoVer , a dataset for evidence extraction and fact verification. It requires models to extract relevant facts from

[1] https://github.com/Kurocaguama/Mu-SHROOM-GIL



*Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1577–1584

multiple Wikipedia articles and determine whether the claim is supported or not supported based on this information. Document retrieval was made by a query-based approach, using cosine similarity between binned uni-gram and bi-gram TF-IDF vectors. For the claim verification model, authors fine-tuned a BERT model to recognize entailment between the claim and the retrieved evidence. A different methodology that focuses on pairwise comparison over entailment is carried out in Vitamin C (Schuster et al., 2021). The authors analyze attention values for tokens in both sentences and signal out contrasting pairs of tokens, this particular work serves as a starting point for our comparison module.

Wei et al. (2024) proposed FEWL, a framework that measures the hallucination score of different LLMs designed for scenarios where the benchmark datasets lack gold-standard answers. Given a set of questions and the correspondent LLMs-generated answers, the framework computes an intermediate truthfulness score weighted by an individual expertise score for each model. Using a set of similar questions, a laziness penalty is applied to the expertise score based on the level of superficialness exhibited by the LLM responses. FEWL has demonstrated effectiveness in mitigating hallucinations by guiding in-context learning and supervised fine-tuning, even in the absence of gold-standard references.

## 3 System overview

Since this task analyzes LLM outputs that aim to answer trivia-like questions, our approach consisted of obtaining a correct answer to the each question in order to analyze differences between the test set answers and our gold standard. The precise evaluation consists of generating two sets of tags called *hard labels* and *soft labels*. Hard labels are character-level intervals that our system considers hallucination, whereas soft labels are probabilities assigned to intervals that correspond to annotator agreement of each interval.

Our system is composed of three modules: question-based information retrieval, text filtering and gold standard generation, comparison between our gold standard and the task's test dataset. The input data corresponds to questions from the test dataset, the output is a fully formatted dataset compatible with the task's evaluator.

### 3.1 Question-based IR

Given a question, a PoST was carried out from which a filter was done to analyze nouns, proper nouns, numbers, and adjectives in the form *nth*; this information was then used as a query to obtain the top $n$ most relevant Wikipedia pages. These $n$ most relevant pages are then passed on to the next section.

### 3.2 Gold Standard Generation

Various gold standards were created, the final submitted model corresponds to the best performing standard on the validation set ($gs_3$). The first gold standard ($gs_1$) is a joined summary of each retrieved page, each summary was obtained using the Wikipedia API's (Jon Goldsmith, February 13th, 2025, Version: 0.8.1) summary function. The second gold standard ($gs_2$) corresponds to an embedding based retrieval of relevant passages within each article's text. Each retrieved page was segmented an embedded using BGE M3 (Chen et al., 2024), these page embeddings were then compared with the corresponding question's embedding and the union of the top 5 most similar segments was considered as the gold standard.

The third and fourth gold standards ($gs_3, gs_4$) were obtained by completing the RAG pipeline using Llama (Llama Team, 2024), and GPT (OpenAI, 2024). Each LLM was tasked to answer a question from the test set based on the retrieved contexts ($gs_1$ or $gs_2$). The respective output was considered the gold standard, $gs_3$ when the response originated from $gs_1$, and $gs_4$ when the response originated from $gs_2$.

### 3.3 Pairwise Comparison

Each gold standard is compared with the test dataset's corresponding answer ($a$), meaning four total comparisons were implemented: $(a, gs_i)$ for $i \in \{1, 2, 3, 4\}$. Rather than opting for a pretrained approach (similar to what was done in Vitamin C), a PoST-based entity comparison was implemented.

Four particular PoST were considered (NOUN, PROPN, NUM, ADJ), as well as the words *yes* or *no* (in English or Spanish). Tagged elements in $s$ but not in $gs_i$ are subsequently analyzed, meaning we shift our focus to the following set: $A = \text{Tags}(a) \setminus \text{Tags}(gs_i)$. For each element in $A$, edit distance was calculated with every element in $gs_i$, elements with edit distance less or equal than 2 are then filtered and tagged as hallucination, particularly as

hard labels. We argue that this tagging corresponds to a hallucination since the gold standard should, in theory, contain a true answer to the question, meaning that differences in nouns and quantities should correspond to incorrect information, and thus hallucinations.

Finally, in order to calculate the soft labels, the words that weren't tagged as hard labels (based on the edit distance cutoff) are tagged with probability $\frac{distance}{10}$. Furthermore, the hard labels are taken and the following calculation is performed in order to determine their soft label probability:

- 1 if the word is a number or a proper noun.
- $1 - \frac{\alpha+1}{2}$ where $\alpha$ is the similarity between the embedding of $gs_i$ and the corresponding word.

### 3.4 System example

An example of our system's working pipe is shown next. Figure 1 is a graphical representation of our system.

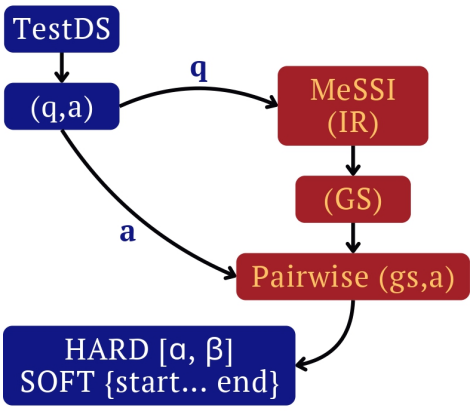


Figure 1: Elements in red are part of our system. $(q, a)$ is an extracted question pair. The interval and the dictionary correspond to our predictions of hard and soft labels.

Consider the following question extracted from task's English test dataset: *In which film does James Bond drive an Aston Martin V12 Vanquish?* The string used for Wikipedia's IR is: *film James Bond Aston Martin V12 Vanquish*. The top 2 most relevant Wikipedia pages are: *Aston Martin Vanquish* and *Aston Martin Vanquish (2012)*.

Gold standards ($gs_1, gs_2, gs_4$) can be found in Appendix A.1. $gs_3$, an LLM generation based on $gs_1$, corresponds to the following answer: *James Bond drove an Aston Martin V12 Vanquish in the 2002 film "Die Another Day"*.

The comparison is then carried out between the gold standards and the model output extracted from the dataset. In this example the test set answer is: *Skyfall*. Table 1 shows the soft and hard labels for the answer and $gs_3$.

| **Label** | $gs_3$ |
|---|---|
| Hard labels | [1,8] |
| Soft labels | {"start":1, "prob":1.0, "end":8} |

Table 1: MESSI's scores

In this particular example the answer is just a single word (tagged as proper noun) that differs from the gold standard. This means that the pairwise comparison will return the whole word as a hard label (as seen in the interval [1,8]), and since the tag is proper noun the soft label is also 1 in the same interval.

## 4 Experimental setup

The experiments were carried out in Python using various packages for each module of the system. For further reference please review our system's repository in Github.

### 4.1 Questions and IR

The input questions were obtained directly from the Mu-SHROOM dataset, based on each question the PoST was done using spaCy for the corresponding language, as well as self-defined functions. Given the processed question, the top 2 most relevant Wikipedia pages were retrieved, this limit was set due to the Wikipedia API rate limits and subsequent LLM token limits. Embeddings where obtained using standard BGE M3 parameters.

### 4.2 LLMs

For the LLM-based gold standard generation, Llama's `Llama-3.2-3B-Instruct`, and GPT's `gpt-4o-mini` checkpoints were used. Both models and checkpoints were selected since they are the closest to state-of-the-art releases and due to resource limitations. Experiments with newer (`Llama-3.3` or `GPT-4o`) and heavier (`Llama-3.2-90B`) models doubled or tripled time during training, taking over 3-4 days of extraction in certain cases. Mistral checkpoints were considered but didn't make the final selection since

the task's dataset contains mainly generations from instruct variants of Mistral models.

Prompt engineering was not carried out for this alternative summary generation. The used prompt can be seen in this paper's appendix. In GPT's case, temperature was set to 0.1. Llama's responses were used over GPT's.

### 4.3 Segment Extraction

The final comparison is based on each gold standard's PoST. In order to maintain consistency throughout the system, spaCy was also used in this section. Calculations described in section 3.3 were implemented without any additional library.

## 5 Results and Analysis

Full results can be found in Mu-SHROOM's official site[2]. The main scores ranked participants by IoU rather than Cor, however an alternative ranking is also available in the official site.

### 5.1 Results

Our system's results achieved an Intersection over Union (IoU) score of 0.4607 in English, and 0.2807 in Spanish. Probability correlation (Cor) scores were 0.5015 and 0.3243 in each respective language. Table 2 shows results in English compared to the best performing team and the task's baseline, Table 3 is analogous but for Spanish. Table 4 shows the amount of correctly extracted Wikipedia pages from the question-based IR module.

| Rank | Team Name | IoU | Cor |
|---|---|---|---|
| 1 | iai_MSU | 0.6509 | 0.6294 |
| 19 | GIL-IIMAS UNAM | **0.4607** | **0.5015** |
| 44 | Baseline (neural) | 0.0310 | 0.1190 |

Table 2: English results

| Rank | Team Name | IoU | Cor |
|---|---|---|---|
| 1 | ATLANTIS | 0.5311 | 0.0132 |
| 18 | GIL-IIMAS UNAM | **0.2807** | **0.3243** |
| 34 | Baseline (neural) | 0.0724 | 0.0359 |

Table 3: Spanish results

### 5.2 Analysis

In both languages our system achieves better results for *Cor* over *IoU*. This probability value corresponds to the Spearman correlation between our system's predicted soft labels and the test set's soft labels, meaning that the correlation is calculated over intervals with a softer cutoff in values and isn't hindered as much by an incorrect prediction unlike with hard labels.

| Test set questions | Correct retrieval | No pages found |
|---|---|---|
| 154 (*en*) | 120 | 2 |
| 152 (*es*) | 84 | 13 |

Table 4: Module 1's retrieval performance

Our system tags intervals with 0 in cases where factual information isn't being discussed since it analyzes only nouns, proper nouns, adjectives and quantities. Since Mu-SHROOM's dataset questions mainly ask about factual information, our 0-tagged intervals tend to coincide with the task annotator's 0-tagged intervals. In addition, the non-zero tags correspond to a value that varies based on the edit distance between words, meaning that the tagging isn't binary and thus has a more lenient cutoff. Regardless, our *Cor* scores don't completely overshadow the *IoU* metric, suggesting that the hard labels predicted by this system could be improved by some factor based on the edit distance between mismatched words.

However our system's main setback comes from the information retrieval and gold standard generation section rather than the pairwise comparison. Before carrying out this comparison, a dataset containing the gold standards is created and used as reference for the following module. Table 4 shows that the retrieval process has room for improvement. In English 77% of retrieved pages are relevant for question answering, whereas in Spanish only 53% of the retrieved pages are relevant. Even in certain cases not a single Wikipedia page is retrieved.

By looking closer into the datasets we can observe various flaws in our pipeline. The Wikipedia based gold standards are inefficient due the sheer length of the final text used as a gold standard. English summaries average 262 words while Spanish ones average 287, on the other hand LLM-based gold standards average 16 and 25 words for each respective language. Considering that the test set's average answer lengths are 39 and 75 respectively, smaller gold standards work better with our pairwise comparison and avoid evaluating differences between a higher amount of words.

Furthermore, looking into $gs_3$'s (Llama gener-

[2] https://helsinki-nlp.github.io/shroom/iou_rankings

ated gold standards based on Wikipedia summaries) dataset[3] we observe that several gold standards aren't actually correct answers, but rather Llama answering that the retrieved text doesn't provide information that actually helps in terms of answering the question. Some of these answers are in the form:

- `Apology` + *The given text doesn't have relevant information.*
- *The given text doesn't have information regarding* ***subject "x"****. However I can give you information regrading* ***subject "x"****.*

From the English test set 50 out of 154 gold standards consisted of these type of answers, in Spanish it was a total of 49 out of 152 (it's worth noting that the second type of answers were considerably more common in the Spanish implementation over the English one). This means that our model carries out an actual pairwise comparison in only two thirds of the cases, lowering performance even before the comparison. In addition to this, the actual LLM-based gold standards aren't always true gold standards. Even in cases were adequate information was retrieved, the augmented generation turned out to be incorrect, further decreasing our system's performance.

This highlights various short comings in our system. Working only with Wikipedia as a database limits our available information to each language's resources in the site. Furthermore automatic summaries are susceptible of losing particular information that can be relevant in cases of very precise questions, regardless if the summary is made by an API or an LLM.

## 6 Conclusion

In this paper we described MeSSI, our pairwise sentence comparison system based on lexical differences, as well as MeSSI's performance and limitations in the SemEval task Mu-SHROOM. In ideal cases our system correctly extracts relevant Wikipedia articles, generates an adequate gold standard and identifies differences in words between gold standards and test set questions.

However, ideal cases aren't always the norm. Wikipedia resources heavily depend on the language, almost 7 million articles are available in English compared with the 2 million in Spanish (Wikipedia, 2025), and Wikipedia unfortunately doesn't contain the whole of humanity's knowledge. On top of this, RAG pipelines are still prone to inadequate text generation, leading to incorrect gold standards used for pairwise comparison.

Regardless our system proved to be competitive in both languages and managed to tag soft and hard labels in a way to keep both values correlated, something that even the best performing Spanish model failed to do.

For future work and possible deployment in a context outside tasks, this system could be improved throughout the pipeline. Initially by enlarging the retrieval database using specialized corpora, textbooks, or general purpose datasets. This way our system is less dependent on one single resource and can cover a wider variety of subjects. The gold standard generation module could be improved by a finer segmentation and embedding of the retrieved documents, this would reduce size of extracted documents and benefit the LLM based generation as well as the pairwise comparison module. Finer text normalization would also improve the system, considering that various Wikipedia based texts contain separators like section and subsection names, or characters like "==".

Finally, alternatives for the comparison module could include a semantic analysis of answers in order to understand which elements coincide to factual information and over just tagged elements, while keeping interpretability of the model. A Transformer based approach, similar to the one presented in Vitamin C, could improve results and focus on token attention over PoST as well as including contextual representation of texts previous to comparison.

## Acknowledgments

This research was funded by UNAM, PAPIIT projects IG400325 and IN104424. Francisco Lopez-Ponce thanks the CONAHCYT scholarship project (CVU: 2045472). Karla Salas-Jimenez thanks the CONAHCYT scholarship project (CVU: 1291359).

## References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity

[3] https://github.com/Kurocaguama/Mu-SHROOM-GIL/blob/main/Datasets/full_pipeline_datasets/en_llm.csv

text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Jon Goldsmith. February 13th, 2025, Version: 0.8.1. Wikipedia: A python wrapper for the wikipedia api.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

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.

Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu. 2024. Measuring and reducing llm hallucination without gold-standard answers. *Preprint*, arXiv:2402.10412.

Wikipedia. 2025. List of wikipedias.

# A Appendix

## A.1 Gold Standards

$gs_1$: *The Aston Martin Vanquish is a grand tourer introduced by British luxury automobile manufacturer Aston Martin in 2001 as a successor to the Aston Martin Virage (1993). The Aston Martin V12 Vanquish, designed by Ian Callum and unveiled at the 2001 Geneva Motor Show, was produced from 2001 to 2007 as the flagship of the marque. A concept car, known as "Project Vantage" and the first Aston Martin design wholly designed by Callum, was built to display the company's vision for a future sports car that could represent Aston Martin's aspirations after the discontinuation of the Virage-based Vantage. The concept car evolved directly into the V12 Vanquish, and featured an advanced carbon fibre and alloy structure, Aston Martin's most powerful V12 engine, and host of new technologies. A specially modified V12 Vanquish was driven by James Bond in the 2002 film Die Another Day. In 2004, a mildly updated version of the first-generation model named "V12 Vanquish S" was introduced featuring a more highly tuned engine and more track-oriented ride and handling. The V12 Vanquish was indirectly replaced by the DBS after 2007. The second-generation "Vanquish" was introduced in 2012, this time based on Aston Martin's existing VH platform – similar to the one that underpinned the DB9. Designed by Marek Reichman and made in the Gaydon facility, the VH platform Vanquish was designed to fill the shoes of the discontinued DBS. In 2017, a "Vanquish S" with a more powerful engine and improved aerodynamics was launched. The second-generation Gaydon Vanquish was succeeded by the DBS Superleggera in 2018. In September 2024, Aston Martin announced the third-generation Vanquish as the successor of the DBS Superleggera.–The second generation of the Aston Martin Vanquish, a grand touring car, was produced between 2012 and 2018 by the British carmaker Aston Martin. It succeeded the DBS, resurrected the name of the 2001–2007 model, and was available as both a coupe and a convertible, the latter known as the Volante. Designed by Marek Reichman, a concept car called the Project AM310 was unveiled at the 2012 edition of the Concorso d'Eleganza Villa d'Este in Lombardy, Italy. The production version was showcased at several events in 2012: a sneak preview at the Goodwood Festival of Speed in July, a presentation to a group of guests at the London Film Museum also in July, and an appearance at the Monterey Car Week in August. The Vanquish, which is based upon the DB9's architecture, namely the vertical/horizontal platform, extensively incorporates aluminium throughout its construction. The Vanquish was produced in Gaydon, a village in Warwickshire, England. Aston Martin unveiled the Vanquish Volante at the 2013 Pebble Beach Concours d'Elegance, with deliveries starting in late 2013. In 2014, the company implemented minor modifications to the Vanquish's engine performance. A more significantly modified version, called the Vanquish S, was launched in 2016; its Volante version was released the following year. The Vanquish S introduced such updates as increased horsepower and torque, and a new body kit. Aston Martin produced the Vanquish Zagato—a special edition—in various body styles, in-*

*cluding a coupe, convertible, shooting brake, and a roadster, the latter dubbed the Speedster. –The Aston Martin DBS is a grand tourer based on the DB9 and manufactured by the British luxury automobile manufacturer Aston Martin. Aston Martin has used the DBS name once before on their 1967–72 grand tourer coupé. The modern car replaced the 2004 Vanquish S as the flagship of the marque. The DBS ended production in 2012 and was succeeded by the second-generation of the Vanquish.*

$gs_2$: *'—The interior of the DBS is a blend of carbon fibre, Alcantara, leather, wood, stainless steel and aluminium surfaces, depending on the buyerś specified options. The door panels are capped with carbon fibre or leather, and utilise carbon fibre door pulls. The fascia is, as standard, matrix alloy and iridium silver centre console or, as an optional extra, piano black fascia and centre console. To achieve even greater weight savings, the carpet has a special lightweight carbon weave. The car is started by means of the "Emotion Control Unit", which was initially developed especially for the DBS but became available for the DB9 and the V8 Vantage as well. The key is made from stainless steel and glass and is inserted into a special slot in the dashboard.== Film appearances ==The DBS was first seen in the 2006 James Bond film Casino Royale, the first film in which Bond was played by Daniel Craig, as a result of Eon Productionś desire to tie the new Bond actor to the franchise heritage with Aston Martin. The only in-car gadget featured in the film is a glovebox/safe that contains a spare pistol, silencer, and a medical kit with a defibrillator in its compartment. Bond uses the car to go to Casino Royale in Montenegro so he can find Le Chiffre. The car is later destroyed when James Bond swerves to avoid hitting Vesper, who had been used as a bait by Le Chiffre to lure Bond after being kidnapped. The cars used in the production were actually prototypes, based on DB9 test vehicles, as the film was produced well before the DBS entered production.The DBS returned for the pre-credits car chase around Lake Garda in the 2008 Bond film Quantum of Solace. The vehicle is colored a dark metallic dark grey, referred to as "Quantum Silver" in Aston Martinś options list, and doesnt́ have any gadgets. In the film, Bond uses the vehicle to deliver Mr. White to M while trying to avoid his pursuers, but is later damaged as a result. The cars used in the film are 2009 model year, German-market spec DBS production vehicles.== Naming confusion ==Some confusion over the name of the production version occurred when some test mules running around the Nürburgring were given DBRS9 badges. However, it would seem that this was only a trick played by the company to confuse spy photographers. The official name of the vehicle was declared to be DBS.== Production ==The DBS was built in Gaydon, Warwickshire. Its engine was built at the Aston Martin engine plant in Cologne, Germany.Production of the DBS totaled 3,381 units, including 2,536 coupes and 845 Volante versions.== References —The DBS was officially unveiled at the 2007 Pebble Beach Concours dÉlegance on 16 August 2007, which featured a brand new exterior colour (graphite grey with a blue tint) which has been dubbed "Lightning Silver".Deliveries of the DBS began in the first quarter of 2008.—The DBS Volante Dragon 88 honours the Year of the Dragon in China. It has 24-carat gold plate on nickel-coated Aston Martin wing badges affixed to the bonnet and rear of each car, bright finish front grille, bonnet meshes and side strakes; choice of three unique 10-spoke designs lightweight forged wheels with a special silver finish, Black brake callipers, a choice of 3 body colours (Amethyst Red, Volcano Red, Champagne Gold) with matching interior upholstery colours (Spicy Red, Deep Purple, Chancellor Red interior for Amethyst Red, Volcano Red, Champagne Gold body respectively), Sahara Tan thread stitching, Piano Black facia trim with a unique gold inlay pattern at dashboard, glass switchgear, headrest embroidery design with rendered using four thread colours (Metallic Gold, Cream Truffle, Winter Wheat and Kestrel Tan) inspired by the Nine-Dragon Wall in Beihai Park, unique laser-etched sill plaques bearing the number and designation, Presentation Box wrapped in the same leather as the interior of their car and lined with Ivory Alcantara (box lid bears an embroidered dragon, with Aston Martin wings embossed on to the front of the lid and a replica sill plaque on the inside of the lid; each box contains an Ownerś Guide with gold detailing, two glass ECUs with leather pouches, a pair of customised Bang & Olufsen earphones with laser-etched Aston Martin wings in a leather pouch) and a 1,000-Watt Bang & Olufsen audio system.The DBS Volante Dragon 88 was unveiled at the 2012 Beijing International Automotive Exhibition and production was limited to 88 units.'*

$gs_4$: *James Bond drives an Aston Martin DBS in the films "Casino Royale" (2006) and "Quantum of Solace" (2008)*

### A.2 Prompts

In both cases the variable *ques* corresponds to whatever question needs to answer, and *context* corresponds to $gs_i$ for $i \in \{1, 2, 3, 4\}$.

**Llama prompt**: *f"You are a bot that answers trivia questions. Be brief, answer in short sentences highlighting important information. If the given text doesn't answer the question, answer as truthfully as you can with your own information. This is the trivia question you need to answer: {ques} This is the text that you should use: {context}"*

**GPT prompt**:

- *system: You are a helpful assistant.*
- *prompt: You are a bot that answers trivia questions. Be brief, answer in short sentences highlighting important information. If the following text doesn't answer the question, answer as truthfully as you can: This is the trivia question you need to answer: {ques}. This is the text that you should use to answer the question: {context}.*

### A.3 Inconsistent Gold Standards

$gs_4$ in Spanish:

- *Lo siento, pero no tengo información sobre un equipo de fútbol argentino llamado "Argenitna" y no puedo encontrar ninguna noticia sobre un capitán llamado "Lionel Messi" en un equipo con ese nombre. Sin embargo, puedo decirte que Lionel Messi ganó la Copa del Mundo con la selección de fútbol de Argentina en 2022, liderada por Lionel Scaloni, no por él mismo.*
- *Lo siento, no tengo información sobre la participación de Chun Jung-myung en una serie de televisión. Sin embargo, puedo decirte que Chun Jung-myung participó en la serie "Absolute Boyfriend" en 2019.*