Leveraging Large Language Models (LLMs) as Domain Experts in a Validation Process

Carlos Badenes-Olmedo[1,†], Esteban García-Cuesta[2,*,†], Alejandro Sánchez-González[2] and Oscar Corcho[2]

[1]Ontology Engineering Group, Departamento de Sistemas Informáticos, Universidad Politécnica de Madrid
[2]Ontology Engineering Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid

Abstract

The explosion of information requires robust methods to validate knowledge claims. On the other hand, there is also an increase interest on understanding and creating methods that helps on the interpretation of machine learning models. Both approaches converge on the necessity of a validation step that clarifies or helps end-users to better understand if the decision or information provided by the model is what is needed or if there is some mismatch between what the artificial intelligent system is suggesting and reality. Large Language Models (LLMs), with their ability to process and synthesize vast amounts of text data, have emerged as potential tools for this purpose. This study explores the utility of LLMs in hypothesis validation in two different scenarios. The first relies on hypothesis generated from explanations obtained by XAI methods or by inherently explainable models. We propose a method to transform the inferences provided by a machine learning model into explanations in natural language, hence linking the symbolic and sub-symbolic areas. The second relies on hypothesis generated with techniques that automatically extract answers from text. The results show that LLMs can complement other XAI techniques and although all LLMs analyzed are able to provide truthfulness-related answers, not all are equally successful.

Keywords

LLMs, knowledge validation, explainable artificial intelligence

1. Introduction

In recent years, the field of artificial intelligence has witnessed a remarkable evolution in natural language processing capabilities, largely driven by the advent of Large Language Models (LLMs). The essence of utilizing these models for Artificial Intelligence (AI) tasks such as knowledge and hypothesis validation lies in their ability to understand, generate, and manipulate human language. This ability is crucial for tasks that require a deep understanding of context and nuances of human communication.

Applying LLMs to real-world scenarios inevitably leads to language generation deviating from known facts (aka "factual hallucination" [1] due to multiple causes (e.g. it may be over-estimating due to overfitting on biased prompts (framing effect)). Several studies have tried to measure these effects but it is still difficult to generalize them outside the specific context where the studies have been performed. In [2] the authors study prompt framing and in-context interference effects showing that large language models are subject to the influences of various hallucinations-inducing causes. This is true for Word Prediction (WP), Question-Answer (QA), and Fact-Checking (FC). Other studies [3] compare the results obtained by machine learning models with those produced by LLMs for diagnostic decision support systems. They propose a processing pipeline for interacting with language models concluding that LLMs models often are ambiguous and provide incorrect diagnoses, being the prompt engineering a critical step in the process. Thus, claim verification has emerged as a key-point to discern between misinformation and real facts. Most of these works [4][5] rely on human-annotated datasets to verify the explanations or decisions but that information is not accessible in hypothesis testing scenarios.

In this paper, we explore the innovative application of Large Language Models (LLMs) as validation tools in fact-checking and hypothesis fact-checking scenarios. Traditionally, the validation of conclusions drawn from data and models in specialized domains has been a task reserved for human experts, largely due to the complexity and domain-specific nature of the required knowledge. However, with the evolution of LLMs, the question arises whether these powerful natural language processing tools can assume a role similar to human experts in the validation of domain-specific knowledge.

The primary goal of this work is to analyze and assess the ability of LLMs to serve as domain experts in a clinical scenario, specifically in validating explanations

SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.
*Corresponding author.
†These authors contributed equally.
carlos.badenes@upmm.es (C. Badenes-Olmedo); esteban.garcia@upm.es (E. García-Cuesta); alejandro.sanchezg@alumnos.upm.es (A. Sánchez-González); oscar.corcho@upm.es (O. Corcho)
0000-0002-2753-9917 (C. Badenes-Olmedo); 0000-0002-1215-3333 (E. García-Cuesta); 0000-0002-9260-0753 (O. Corcho)

(and thereof validating the models) derived from classical machine learning decision models. These explanations, presented in the form of affirmative statements such as "the hypertension increase the risk of death from COVID19", are transformed into questions (for example, "Does hypertension mean an increased risk of death from COVID-19?") to be presented to the LLMs. This approach allows us to directly evaluate the LLM models' as a knowledge base to validate specific claims within the domain, offering a unique perspective on their applicability as validation tools in scientific and clinical contexts.

This work aims to address the following research questions: RQ1) What effect does the variation in the number of options within a fact-checking question have on the responses provided by Large Language Models (LLMs)? RQ2) How consistent are the boolean answers (i.e. yes, or no) provided by Large Language Models (LLMs)? RQ3) What is the impact of integrating machine learning inferences with Large Language Models (LLMs) on enriching and validating the explanations?

Through this analysis, we seek not only to understand the level of knowledge and accuracy of LLMs in specialized domains but also to investigate their potential to complement or, in some cases, replace the need for human peer review in the validation stage of scientific conclusions.

Our contributions are:

1. a novel assessment method that integrates machine learning inferences with Large Language Models (LLMs) to generate fact-checking (FC) type questions.
2. a study on the variability and consistency of responses provided by LLMs in multiple-choice questions and scenarios with established ground truths..
3. an investigation into the variability of explanations provided by LLMs in scenarios involving fact-checking (including questions with multiple factual options) and fact recovery, offering a comprehensive understanding of LLMs' explanatory capabilities and their potential for enhancing AI interpretability.

2. Related works

Prompt framing effect The study of the prompt framing effect reveals that the performance of Large Language Models (LLMs) is highly dependent on the construction of the prompts, with a significant focus on the consistency of LLMs' responses to similar prompts. This concept, discussed in [6], [7], and [8], examines LLMs' ability to provide consistent outputs for semantically similar prompts and their sensitivity to hallucination-inducing inputs. The examination of LLMs under different conditions, such as varying the context and structure of the prompts, sheds light on their performance variability and the strategies for optimizing accuracy.

Building on this foundation, the interplay between context, choice structure, and decision-making, as explored in [9], [10], [11], and [12], directly relates to the challenges LLMs face. This parallel between human and computational decision-making processes emphasizes the importance of carefully designed prompts and the strategic manipulation of choice options to improve LLM reliability and decision accuracy. Through innovative decision-making strategies and prompt engineering techniques proposed in [13], [14], [15], [16], and [17], the nuanced approach to prompt framing is critical for enhancing LLM interactions and understanding. This body of work collectively illustrates a key insight: adjusting the number of options and the framing of prompts can profoundly influence the effectiveness of LLMs in verifying statements and making decisions, bridging the gap between consistency in output and the complexity of input conditions.

Explainable AI and LLMs Interpretability and explainability in Machine Learning (ML) refers to the ability to make understandable an ML model's workings. This is particularly vital in high-risk applications and desirable in most cases. The burgeoning field of research that addresses to foster this ability is known as eXplainable Artificial Intelligence (XAI). A variety of XAI methods have been developed in recent years. They may be related to intrinsically interpretable models or to "black box" models, but all pursue coherent and meaningful explanations for the audience. As an example, SHAP (SHapley Additive exPlanations) is one of the most widely used XAI model agnostic techniques. It is based on concepts from game theory that allow the computing, which are the features that contribute the most to the outcomes of the black-box model, by trying different feature set permutations [18]. LIME (Local Interpretable Model-agnostic Explanations) is another well known example that builds a simple linear surrogate model to explain each of the predictions of the learned black-box model [19]. There are also some interpretable ML models such as logistic regression, Generalised Linear Models (GLMs), or Generalised Additive Models (GAMs).There are some attempts to facilitate the comprehension of some XAI methods providing new tools to end-users. At [20] a new GPT x-[plAIn] is proposed to transform the output explanations provided by those methods (e.g. SHAP or LIME) to natural language that contains the technical descriptions of the results. Despite the improvements in end-user satisfaction, this work does not include any enrichment or additional information that could contextualize not only the explanations themselves, but also the meaning and

validation of the application domain. In [21] the authors propose to use LLMs to facilitate decision-making processs by the end users providing concise summaries of varios XAI methods tailored for different audiences. This can be viewed as LLM enhanced XAI explainer trying to bridge the gap between complex AI technologies and their practical applications.

Veracity and truth extraction The exploration of truth within the realm of big data and its verification through LLMs embodies a complex interaction between technological advancements and the multifaceted nature of truth. The assembly method, as proposed by [22], marks a significant step in addressing the challenge of data veracity by combining individual truth discovery methods to mitigate the effects of limited labeled ground truth availability. This approach lays the groundwork for further research on the role of technology in differentiating between truth and falsehood. Furthermore, research on linguistic indicators of truth and deception, such as that of [23], reveals the potential of linguistic complexities and immediacy to act as markers to distinguish between truthful and deceptive narratives, enriching the conversation about truth verification in digital communications.

Recent advances in artificial intelligence, notably the conceptualization of models such as InstructGPT as "Truth Machines" by [24], highlight ongoing efforts to define and operationalize truth through sophisticated data analysis and model architectures. Currently, innovative methodologies such as the *DoLa* decoding strategy by [25] and the development of truthfulness *personas* by [26] aim to enhance the factuality and reliability of LLM outputs. These strategies not only address the challenge of hallucinations in model responses but also open up new pathways for embedding truthfulness within AI systems, underscoring the dynamic nature of research focused on achieving reliable knowledge verification and decision-making processes in the digital era.

3. Approach and Problem Setup

Our proposal involves using LLMs as knowledge bases to evaluate the outcomes of machine learning models by answering Boolean questions derived from the models' inferences. This approach aims to harness the comprehensive knowledge and understanding capabilities of LLMs to verify the accuracy and reliability of inferences made by machine learning models, thereby providing a novel method for validating AI-generated insights through direct, yes-or-no questioning.

3.1. Large Language Models

A range of LLMs have been developed in the last years. GPT-4, developed by OpenAI, is a state-of-the-art LLM known for its deep learning architecture. As part of the Generative Pre-trained Transformer series, it includes a large network of multi-layer transformers, capable of processing sequential data and preserving textual dependencies in the long term. This version marks a significant advancement over its predecessors by scaling up the number of parameters and broadening the diversity of its training data, thus enhancing its ability to generate coherent and contextually relevant text based on the input it receives [27].

Moreover, Google's DeepMind project Gemini, is a key competitor to GPT-4. Gemini is a family of models built on top of transformer decoders that employ attention mechanisms, analogous to GPT-4. Gemini Pro, the second model in the family in terms of size, has been optimized for both cost and latency, offering considerable performance improvements across numerous tasks; it is designed to understand, reason, and generate outputs across various types of data, including text [28].

Similarly, Llama 2 constitutes a collection of pretrained and fine-tuned LLMs that is distinctive from the models mentioned due to its open-source nature [29]. This group of models developed by Meta includes two models (Llama 2 and Llama 2-Chat) with different versions that adjust the number of parameters: 7B, 13B and 70B.

Mistral represents another significant collection of LLMs, characterized by their advanced reasoning capabilities and a robust performance. Their largest model, Mistral Large, demonstrates state-of-the-art results across a variety of benchmarks, including areas such as common sense, reasoning, and knowledge-based tasks [30]. The Mistral family also includes open-source models that surpass certain versions of Llama 2 in several benchmarks, as documented by [31].

3.2. Datasets

Covid19 explanations The questions included in Table 1 are created from a clinical study [32]. In that study one thousand and three hundred thirty-one COVID-19 patients (medium age 66.9 years old; males n= 841, medium length of hospital stayed 8 days, non-survivors n=233) were analyzed. Based on the hypotheses raised in the study, the questions are constructed. Questions Q2, Q3, Q4, Q5, Q6, Q7, and Q8 were identified as significant using a regression Cox model and Q1, Q9, Q10 were identified as significant by univariate analysis. Q1 was also identified as 1 of the most important variables using SHAP explanations over LSTM learned model using the same Covid19 dataset. By domain knowledge and based on model explanations we can set Q1, Q2, Q3, Q4, Q5, Q6,

Table 1
Consistency questions dataset

	Consistency
Q1	Does hypertension mean an increased risk of death from COVID-19?
Q2	Does a low platelet count mean an increased risk of death from COVID-19?
Q3	Does a high leukocyte count at emergency mean an increased risk of death from COVID-19?
Q4	Does older age mean an increased risk of death from COVID-19?
Q5	Does male gender mean an increased risk of death from COVID-19?
Q6	Does previous chronic therapy with steroids mean an increased risk of death from COVID-19?
Q7	Does not treating with hydroxychloroquine mean an increased risk of death from COVID-19?
Q8	Does oxygen saturation at emergency mean an increased risk of death from COVID-19?
Q9	Does no early prescription of lopinavir/ritonavir mean an increased risk of death from COVID-19?
Q10	Does no treatment with steroid bolus mean an increased risk of death from COVID-19?

and Q8 as positive truth answers. We did not include Q7 as a positive response (but controversy), despite being obtained by the Cox model explanations, because there was controversy about the use of hydroxychloroquine during the pandemic and although it was initially considered as a drug to reduce the risk of mortality, it was later contradicted by other studies and was not recommended by the World Health Organization. Therefore, the variables that were obtained only by the univariate analysis (Q9 and Q10) are proposed as controversy answers.

It is important to highlight that all the questions adhere to a consistent structure to optimize the performance of the LLM. Specifically, each question is framed as "*Does #hypothesis# mean an increased risk of death from COVID-19?*". This uniformity ensures that the LLM's responses are directly comparable and minimizes variability that could arise from differing question formats. It also allows to test hypothesis obtained by the explainability models.

Veracity dataset The Stanford Question Answering Dataset (SQuAD) [33] has been extensively used in the scientific literature for the development of Question Answering (QA) language models, serving as a benchmark to assess the abilities of these models in understanding and processing natural language queries. As a rich compilation of questions and answers based on Wikipedia articles, SQuAD challenges models to provide accurate answers by comprehending the context provided in the passages.

In our work, we retrieved a subset of questions from the SQuAD dataset to specifically validate the knowledge conveyed by LLMs. This targeted evaluation was designed to determine the precision of the LLM answers compared to the gold standard answers of the data set. This method of validation not only tests the LLMs' understanding of complex texts, but also assesses their reliability in providing information that matches human-curated answers.

3.3. Use Cases

Three use cases (UC) have been designed to address previous research questions, focusing on the practical applications and implications of using LLMs to validate machine learning inferences. The first area investigates the influence of varying the number of options in fact-check questions on LLM responses, aiming to understand how choice diversity impacts LLM accuracy. The second focuses on assessing the consistency of boolean (yes or no) answers provided by LLMs, evaluating their reliability in delivering steady responses. Lastly, we explore the effects of combine machine learning inferences with LLMs to both enrich and validate the explanations of these models. This last use case uses the Covid19 dataset to create a ML model and the SHAP technique to obtain a set of important features that later are enriched with LLMs.

The models used in this study include "gpt-4" from OpenAI, "mistral-large-2402" from Mistral AI, "gemini-1.0-pro-001" from Google, and "llama-2-70b-chat" from Meta AI. In addition, the temperature parameter was set to the lowest possible value to ensure the most deterministic behavior in the LLMs. Temperature controls the randomness of the generated output, with a lower value leading to more deterministic outputs by favoring the most likely predictions. Therefore, in most models, the temperature value was set to 0 to minimize randomness. However, it is important to note that for the Llama 2 model, the minimum supported temperature value is 0.01. Despite this slight deviation from 0, the aim remains the same: to achieve the lowest possible level of randomness in the output.

UC1: Fact Density Impact Analysis It examines the performance of LLMs in delivering binary responses (*"yes"* or *"no"*) versus incorporating a third option (*"controversy"*) to introduce an element of uncertainty. This evaluation aims to measure the models' performance in terms of *veracity*, exploring how the structure of the response options affects the LLMs' ability to provide accurate and reliable answers in fact-checking scenarios.

Table 2 presents the prompts used in three scenarios to evaluate *veracity*, allowing the model to use binary responses or multiple options, and requesting the model to act as an expert in the clinical domain, providing precise

Table 2
Use Case 1 Contexts

	Prompt
Context 1	You are an expert on COVID-19 and your duty is to answer questions related to the topic only with yes or no followed by the explanation that validates the answer in a maximum of 2 sentences.
Context 2	You are an expert on COVID-19 and your duty is to answer questions related to the topic only with yes, no or controversy followed by the explanation that validates the answer in a maximum of 2 sentences.
Context 3	You are a medical expert and your duty is to answer medical questions in a single sentence in a precise and brief manner.

and concise responses. The use of the parameter *max-tokens* inadvertently caused responses to be abruptly cut, leading to nonsensical outcomes. Consequently, we directed the model within the context to be precise and concise, with the aim of minimizing this issue and enhancing the clarity and relevance of its answers. This additional context of evaluation was designed to gauge the model's capacity to offer accurate and reliable answers when positioned as a domain-specific authority, further enriching our understanding of its performance in delivering veracious responses within specialized scenarios. This distinction allows for a detailed examination of how the inclusion of an "controversy" option alongside traditional "yes" or "no" answers influences the model's response behavior in our Use Case 1 analysis.

UC2: Consistency and Veracity Evaluation Use Case 2 distinguishes between two methods of evaluating LLM consistency based on the availability of ground truth. In the first approach, where the true answer is not available, consistency is assessed by comparing the LLM's responses against each other. This method focuses on the internal consistency of the model's answers. In the second approach, where a known true answer exists, the LLM's responses are evaluated against this ground truth to measure the model's accuracy and reliability in providing consistent and correct answers, a quality referred to as veracity.

On the one hand, the first approach or consistency evaluation aims to assess the stability of responses from LLMs through repeated inquiries. By introducing an algorithm 1 to systematically evaluate consistency within the Covid19 dataset, we probe each question in the dataset multiple times using the question and *Context 1* as the prompt. This method allows us to gauge the LLMs' consistency using the metrics described in Section 3.4. Similarly, the same algorithm is used with *Context 2*.

The following algorithm was deployed twice for each LLM, once for each of the two contexts, and the temperature parameter was minimized to enhance response determinism. This methodology provides a nuanced understanding of the models' consistency by ensuring controlled conditions and leveraging the lowest possible temperature setting to maximize the determinism of the models' responses.

Algorithm 1 Evaluate the consistency of a single LLM

for each question q_i in *dataset1* **do**
  Initialize *Responses* to an empty list
  for $i \leftarrow 1$ **to** 10 **do**
    response $r \leftarrow$ AskLLM(q_i, *context1*)
    Append r to *Responses*
  end for
  SemanticSimilarity $\leftarrow$ CalculateSemanticSimilarity(*Responses*)
  Overlap $\leftarrow$ CalculateOverlap(*Responses*)
  ROUGE $\leftarrow$ CalculateROUGE(*Responses*)
  BLEU $\leftarrow$ CalculateBLEU(*Responses*)
  Store metrics for further analysis
end for

On the other hand, the veracity evaluation involves the use of ground truth. Therefore, akin to the previous method, we employ a different algorithm (see Algorithm 2) designed to assess the veracity of each response from each model. The key difference in this approach is that when invoking the LLM, both the response along with its context (*Context 3*) and the ground truth for each response ("q_i answer") are provided. This enables a direct comparison between the LLM's responses and the known accurate answers.

UC3: XAI Enhancement and Validation Use Case 3 involves leveraging machine learning inferences and LLMs to enrich and validate explanations. We propose to utilize important features identified by XAI techniques, such as SHAP, to augment information and validate explanations. This involves transforming explanations into binary questions that LLMs can answer, with prompts

Algorithm 2 Evaluate the veracity of a single LLM

for each question q_i in *dataset2* **do**
  Initialize *Responses* to an empty list
  for $i \leftarrow 1$ **to** 10 **do**
    response $r_i \leftarrow$ AskLLM(q_i, *context3*)
    Append r_i to *Responses*
  end for
  SemanticSimilarity $\leftarrow$ CalculateSemanticSimilarity(*Responses*, q_i answer)
  Overlap $\leftarrow$ CalculateOverlap(*Responses*, q_i answer)
  ROUGE $\leftarrow$ CalculateROUGE(*Responses*, q_i answer)
  BLEU $\leftarrow$ CalculateBLEU(*Responses*, q_i answer)
  Store metrics for further analysis
end for

that contain both the question and relevant contexts. By constructing queries to directly link significant features with real-world results (e.g. 'Does #hypothesis# mean an increased risk of death from COVID-19?), we bridge the gap between XAI insights and practical applications. Additionally, by instructing LLMs to respond with "yes" or "no" and provide validating explanations, we achieve dual objectives of validating and enriching responses, prompting LLMs to elaborate on pertinent features.

3.4. Metrics

A suite of metrics has been implemented to evaluate the consistency and veracity of the LLMs. This suite includes semantic similarity, token overlap, and the ROUGE and BLEU metrics.

Semantic similarity is a measure of the degree to which two concepts (such as words, phrases, or sentences) are related in terms of their meanings within a given semantic space. In formal terms, semantic similarity can be quantified based on the distance or closeness of the concepts in a multi-dimensional space, where each dimension represents a feature of the concept's meaning. The closer two concepts are in this space, the more semantically similar they are.

Diverse methods for calculating semantic similarity are analysed in [34], encompassing a range of approaches. However, this research will specifically utilize cosine similarity in conjunction with sentence embeddings. We will use Sentence-BERT, a variation of BERT(Devlin et al., 2018) optimized for sentence-level embeddings, due to their proven efficiency [35]. In particular, this research utilizes the "all-MiniLM-L6-v2" model for its remarkable balance between high performance and speed. Despite being one of the smallest models in terms of size, it stands out for its rapid processing capabilities.

Overlap as a metric refers to the method of quantifying similarity based on the common tokens (words or other meaningful elements) that appear in two sentences. This metric is used to assess how much shared content exists between both sentences, indicating their consistency or similarity in terms of the information they convey.

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. ROUGE includes a collection of metrics designed for the formal evaluation of text generation models such us summarization or machine translation. In the evaluation of responses generated by a LLM, the use of the ROUGE metric can be justified by its ability to quantitatively measure the lexical overlap across different responses generated by the LLM itself. This is accomplished utilizing the ROUGE-L variant, which employs the Longest Common Subsequence (LCS) between two sentences as a basis for computing recall, precision, and the F_1 score derived from both [36].

BLEU stands for Bilingual Evaluation Understudy. BLEU is a metric initially conceived for evaluating the quality of text translated by machine translation systems by comparing it with one or more reference translations [37]. Unlike ROUGE, which is recall-oriented, BLEU emphasizes precision. It assesses how many words or phrases in the machine-generated text appear in the reference texts. This metric calculates n-gram (contiguous sequences of n items from a given sample of text) precision for different lengths and combines them through a weighted geometric mean, incorporating a brevity penalty to discourage overly short translations [37].

This precision-oriented approach is particularly valuable when the objective is to ensure that certain key information is consistently represented in the LLM's outputs. We computed the BLEU metric by treating each LLM response as a "translation" and comparing it to other responses. BLEU can highlight the extent to which the LLM is capable of producing responses that contain expected and relevant content. This method offers a complementary perspective to the recall-focused metric ROUGE, providing a balanced assessment of the LLM's performance.

4. Evaluation and Results

In this section we evaluate the different use cases.

4.1. UC1 Results

The first use case focuses on evaluating how the structure of response options presented to the LLM influences the performance of the models' accuracy and reliability. This evaluation was addressed by using different contexts: *Context 1* which employs a binary response (such as "yes"

Table 3
Context 1 vs *Context 2* responses for all LLMs

		GPT		Mistral		Gemini		Llama2	
	Expected	Context 1	Context 2	Context 1	Context 2	Context 1	Context 2	Context 1	Context 2
Q1	yes	yes	yes	yes	yes	yes	controversy	yes	yes
Q2	yes	yes	yes	yes	yes	yes	controversy	no	controversy
Q3	yes	yes	yes	yes	yes	yes	controversy	yes	controversy
Q4	yes	yes	yes	yes	yes	yes	yes	yes	yes
Q5	yes	yes	yes	yes	yes	yes	yes	no	controversy
Q6	yes	yes	controversy	yes	controversy	no	controversy	yes	controversy
Q7	controversy	no	controversy	no	controversy	no	controversy	no	controversy
Q8	yes	yes	yes	yes	yes	yes	yes	yes	controversy
Q9	controversy	no	controversy	yes	controversy	no	controversy	yes	controversy
Q10	controversy	yes	controversy	yes	controversy	no	controversy	yes	controversy

or "no"), and *Context 2*, which introduces a third element associated to uncertainty characterized as "controversy".

Table 3 shows the results of the models' responses for each question. It is important to clarify that although multiple responses are generated for each question (specifically 10), the table presents only a single value in each cell. This reduction is justified because the answers ("yes", "no" or "controversy") do not vary across iterations. What varies is the model's explanations of the responses, not the answer itself.

However, some differences can be noted both in the responses generated by a LLM with different contexts for the same question and in the performance across various language models (e.g. Q1 in *Context 2* is answered as "yes" by GPT but "controversy" by Gemini). These variations reveal that while some differences can be attributed to the introduction of 'controversy' in response options (e.g. GPT Q7), others may not have such a clear justification (e.g. Gemini Q2).

Optimally, each LLM should make three justified variations (Q7, Q9, and Q10) when introducing the uncertainty option with the second context, due to the limitation of *Context 1* to binary "yes" or "no" answers. For GPT, an analysis of the responses between contexts reveals a mixed outcome: 3 of the variations presented are deemed correct (Q7, Q9, and Q10), indicating that the model accurately handled both contexts. Conversely, the model's responses to the question Q6 is classified as wrong variations, suggesting inaccuracies in dealing with different contexts.

Similarly, the performance of the other models is as follows:

- Mistral accurately handles 3 variations (Q7, Q9, and Q10) but had an error in Q6.
- Gemini stands out by correctly handling 3 variations (Q7, Q9, and Q10) but falls short by producing 4 unjustified incorrect variations (Q1, Q2, Q3, Q6), including a notable discrepancy in Question 6 where the expected answer was "yes", but the output was "no".
- Llama2 demonstrates accuracy in variations of Q7, Q9, and Q10. However, it produces unjustified variations in Q2, Q3, Q5, Q6, and Q8. Furthermore, it provides incorrect answers for Q2 and Q5, where "yes" was expected, but "no" was output.

Our findings suggest that introducing the option of "controversy" as a potential response significantly influences the behavior of the analyzed LLMs, leading to a noticeable shift in their response patterns. Across various models, including GPT and Mistral, where the response changed in 4 out of 10 instances, Gemini with a change in 7 out of 10 instances, and Llama2 showing a change in 8 out of 10 instances, there is a marked preference for selecting "controversy" over a definitive "yes" or "no". This tendency persists irrespective of the model in question and appears to reflect a broader pattern: when presented with the "controversy" option, models consistently avoid negative responses, opting instead to categorize statements as controversial. This behavior suggests a higher level of confidence in asserting conclusions rather than denying them. While for GPT and Mistral, 75% of these shifts towards "controversy" can be considered justified, enhancing the quality of the output, the justification for this change drops to 43% for Gemini and 37% for Llama2, indicating variability in how these adjustments align with the underlying data uncertainty.

4.2. UC2 Results

In this section, we present the results from the second use case, which are detailed in Tables 4 and 5. These tables show the average performance metrics for consistency and veracity -namely, semantic similarity, overlap, ROUGE and BLEU scores - for each model across various datasets. These metrics were computed for each question within the datasets, with averages provided to give a view of each model's performance under two different contexts (i.e *Context 1* and *Context 2*) for consistency evaluation (Table 4; the consistency results per questions are

Table 4
Average consistency evaluation

		Semantic similarity	Overlap	ROUGE	BLEU
GPT	Context 1	0,983	0,888	0,844	0,770
	Context 2	0,980	0,897	0,853	0,763
Mistral	Context 1	1,000	1,000	1,000	1,000
	Context 2	0,999	0,995	0,992	0,988
Gemini	Context 1	1,000	0,996	0,996	0,992
	Context 2	0,999	0,996	0,993	0,986
Llama2	Context 1	1,000	0,998	0,997	0,996
	Context 2	0,999	0,991	0,989	0,983

Table 5
Average veracity evaluation

	Semantic similarity	Overlap	ROUGE	BLEU
GPT	0,740	0,464	0,301	0,408
Mistral	0,727	0,403	0,222	0,380
Gemini	0,676	0,415	0,273	0,328
Llama2	0,734	0,466	0,239	0,393

Table 6
Average consistency of explanations

	Semantic similarity	Overlap	ROUGE	BLEU
GPT	0,916	0,659	0,541	0,375
Mistral	0,931	0,673	0,570	0,389
Gemini	0,915	0,600	0,442	0,247
Llama2	0,905	0,552	0,411	0,219

provided at Appendix TableA1) and a third context (i.e. *Context 3*) for veracity evaluation (Table 5; the veracity results per questions are provided at Appendix TableA3).

Our analysis reveals no significant difference in performance between the first two contexts evaluated for consistency, where all LLMs demonstrated high levels of consistency. Mistral achieved perfect consistency scores, while Gemini and Llama2 were nearly perfect. However, GPT showed the lowest consistency (for all metrics including semantic similarity), even with the temperature parameter set to the lowest level, indicating potential variability in its response generation process.

When comparing the models' performance to the ground truth data for veracity (see Section 3.2), GPT stands out by achieving the best results across all metrics, indicating that its responses, on average, align more closely with the ground truth than those of the other models. Llama2 follows closely behind as the second-best performer, with Gemini and Mistral trailings and their positions varying depending on the metric applied. These findings suggest that while GPT may struggle with consistency relative to its peers, it excels in generating responses that are more closely aligned with verifiable facts, highlighting a nuanced trade-off between consistency and veracity across different LLMs.

4.3. UC3 Results

It examines the use of prompts that transform explanations into binary questions that contain both the question and relevant contexts related with the fact-checking under analysis (e.g. *Context 1* 'You are an expert on COVID-19 and your duty is to answer questions related to the topic only with yes or no followed by the explanation that validates the answer in a maximum of 2 sentences.'). Table 7 shows example of responses for Q1, Q2, Q3 from GPT-4. Q1 enriches the fact-checking response adding information related with the consequences of having hypertension and how they are related to higher death risk. Q2 enriches the response adding reasons why the important feature (i.e. platelet) plays a crucial role that may lead to high risk of death. Last, Q3 response enriches the response indicating that a high leukocyte can be a symptom of severe Covid19. At table 8 we studied syntactically the number of words that contain the explanation and also the average number of words per sentence. Llama2 and Mistral have larger explanations and also syntactically are slighly more comples (Llama2 has $\approx$ 28 words per sentence for context2). Gemini provides the shortest explanations and also the lowest syntactic complexity (36.72 number of words average and 19.91 words per sentence). Similarly to previous use cases we analyzed the differences between *Context 1* and *Context 2* explanations (including the controversy as an option in the second) to measure how different are the explanations. According to all metrics the results show that the LLM that change the most is Llama2 (i.e. ROUGE 0.411), followed by Gemini (i.e. ROUGE 0.442), GPT-4 (i.e. ROUGE 0.541) and Mistral (i.e. ROUGE 0.570) (see Table 6 for other metrics).

Table 7
Examples of responses for GPT-4

	Example of responses
Q1	Yes, hypertension has been identified as a risk factor for severe outcomes in patients with COVID-19. Studies have shown that patients with hypertension are more likely to experience severe symptoms or complications, including death, from the virus.
Q2	Yes, studies have shown that a low platelet count, or thrombocytopenia, can be associated with a higher risk of severe disease and mortality in patients with COVID-19. This is because platelets play a crucial role in the body's immune response, and a low count can impair the body's ability to fight off infections.
Q3	Yes, a high leukocyte count, or leukocytosis, can indicate a severe infection or inflammation in the body, including severe COVID-19. Studies have shown that patients with severe COVID-19 often have leukocytosis, which is associated with a higher risk of mortality.

Table 8
Average number of words of explanation per text and per sentence

		Avg words per text	Avg words per sentence
GPT	Context 1	39,38	19,69
	Context 2	37,46	22,65
	Average	38,42	21,17
Mistral	Context 1	53,30	21,62
	Context 2	45,70	20,20
	Average	49,5	20,91
Gemini	Context 1	37,70	18,85
	Context 2	35,74	20,98
	Average	36,72	19,915
Llama2	Context 1	57,70	23,20
	Context 2	50,76	27,69
	Average	54,23	25,445

5. Conclusions

In this paper we studied the effect of variation in the number of options within fact-checking questions, the consistency and truthfulness of the answers, and the capabilities to enrich fact-checking with explanations. We also proposed to link explanations from machine learning models to LLMs by using those explanations to create a fact-checking type input question. We measured coherence and veracity using state-of-the-art metrics such as semantic similarity, overlap, ROUGE and BLEU, and the results show that Mistral is the most coherent LLM. Notably, Gemini and Llama2 obtained similar results and GPT was slightly behind. Furthermore, we conclude that fact-cheking consistency does not depend on the number of options but explanations' consistency does. This is relevant because it means that a different number of options not only may change the fact response but will also be able to justify it differently. Further research should be done to analyze in depth to what extend these differences might even imply contradictory responses. As for the truthfulness analysis, we observed that GPT obtained the best results on average and can be considered quite accurate.

Acknowledgments

This work has been funded by the project "Inteligencia Artificial eXplicable" IAX grant of the Young Researchers 2022/2024 initiative of the Community of Madrid.

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–45.

[2] W. Wang, B. Haddow, A. Birch, W. Peng, Assessing the reliability of large language model knowledge, arXiv:2310.09820 (2023). URL: https://doi.org/10.48550/arXiv.2310.09820].

[3] L. Caruccio, et al., Can chatgpt provide intelligent diagnoses? a comparative study between predictive models and chatgpt to define a new medical diagnostic bot, Expert Systems with Applications 235 (2024) 121186. URL: https://www.sciencedirect.com/science/article/pii/S0957417423016883. doi:https://doi.org/10.1016/j.eswa.2023.121186.

[4] Y. Jin, X. Wang, R. Yang, Y. Sun, W. Wang, H. Liao, X. Xie, Towards fine-grained reasoning for fake news detection, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 5746–5754. URL: https://ojs.aaai.org/index.php/AAAI/article/view/20517. doi:10.1609/aaai.v36i5.20517.

[5] D. Wadden, K. Lo, L. L. Wang, A. Cohan, I. Beltagy, H. Hajishirzi, MultiVerS: Improving scientific

claim verification with weak supervision and full-document context, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 61–76. URL: https://aclanthology.org/2022.findings-naacl.6. doi:10.18653/v1/2022.findings-naacl.6.

[6] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, Y. Goldberg, Measuring and improving consistency in pretrained language models, Transactions of the Association for Computational Linguistics 9 (2021) 1012–1031. URL: https://aclanthology.org/2021.tacl-1.60. doi:10.1162/tacl_a_00410.

[7] H. Raj, V. Gupta, D. Rosati, S. Majumdar, Semantic consistency for assuring reliability of large language models, arXiv preprint arXiv:2308.09138 (2023).

[8] Q. Dong, J. Xu, L. Kong, Z. Sui, L. Li, Statistical knowledge assessment for large language models, Advances in Neural Information Processing Systems 36 (2024).

[9] E. A. Maylor, M. A. Roberts, Similarity and attraction effects in episodic memory judgments, Cognition 105 (2007) 715–723. URL: https://www.sciencedirect.com/science/article/pii/S0010027706002587. doi:https://doi.org/10.1016/j.cognition.2006.12.002.

[10] K. V. Morgan, T. A. Hurly, M. Bateson, L. Asher, S. D. Healy, Context-dependent decisions among options varying in a single dimension, Behavioural Processes 89 (2012) 115–120. URL: https://www.sciencedirect.com/science/article/pii/S0376635711001719. doi:https://doi.org/10.1016/j.beproc.2011.08.017, comparative cognition: Function and mechanism in lab and field.

[11] P. Pezeshkpour, E. Hruschka, Large language models sensitivity to the order of options in multiple-choice questions, ArXiv abs/2308.11483 (2023). doi:10.48550/arXiv.2308.11483.

[12] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2021) 1 – 35. doi:10.1145/3560815.

[13] J. Zhan, H. Jiang, Y. Yao, Three-way multiattribute decision-making based on outranking relations, IEEE Transactions on Fuzzy Systems 29 (2021) 2844–2858. doi:10.1109/tfuzz.2020.3007423.

[14] T. Haladyna, S. Downing, How many options is enough for a multiple-choice test item?, Educational and Psychological Measurement 53 (1993) 1010 – 999. doi:10.1177/0013164493053004013.

[15] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, ArXiv abs/2302.11382 (2023). doi:10.48550/arXiv.2302.11382.

[16] S. Oymak, A. Rawat, M. Soltanolkotabi, C. Thrampoulidis, On the role of attention in prompt-tuning (2023) 26724–26768. doi:10.48550/arXiv.2306.03435.

[17] A. Bhargava, C. Witkowski, M. Shah, M. W. Thomson, What's the magic word? a control theory of llm prompting, ArXiv abs/2310.04444 (2023). doi:10.48550/arXiv.2310.04444.

[18] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[19] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[20] P. Mavrepis, G. Makridis, G. Fatouros, V. Koukos, M. M. Separdani, D. Kyriazis, Xai for all: Can large language models simplify explainable ai?, 2024. arXiv:2401.13110.

[21] P. Mavrepis, G. Makridis, G. Fatouros, V. Koukos, M. M. Separdani, D. Kyriazis, Xai for all: Can large language models simplify explainable ai?, ArXiv abs/2401.13110 (2024). URL: https://api.semanticscholar.org/CorpusID:267199844.

[22] L. Berti-Équille, Data veracity estimation with ensembling truth discovery methods, 2015 IEEE International Conference on Big Data (Big Data) (2015) 2628–2636. doi:10.1109/BigData.2015.7364062.

[23] J. Burgoon, L. Hamel, T. Qin, Predicting veracity from linguistic indicators, Journal of Language and Social Psychology 37 (2012) 603 – 631. doi:10.1177/0261927X18784119.

[24] L. Munn, L. Magee, V. Arora, Truth machines: Synthesizing veracity in ai language models, AI & SOCIETY (2023). doi:10.1007/s00146-023-01756-4.

[25] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. R. Glass, P. He, Dola: Decoding by contrasting layers improves factuality in large language models, in: Learning Representations (ICLR), 2024 International Conference on, volume abs/2309.03883, 2023. doi:10.48550/arXiv.2309.03883.

[26] N. Joshi, J. Rando, A. Saparov, N. Kim, H. He, Personas as a way to model truthfulness in language models, ArXiv abs/2310.18168 (2023). doi:10.48550/arXiv.2310.18168.

[27] OpenAI, J. Achiam, S. Adler, S. Agarwal, Gpt-4 technical report, ArXiv abs/2303.08774 (2023). URL: https://api.semanticscholar.org/CorpusID:257532815.

[28] G. T. G. R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, Gemini: A family of highly capable multimodal models, ArXiv abs/2312.11805 (2023). URL: https://api.semanticscholar.org/CorpusID:266361876.

[29] H. Touvron, L. Martin, K. R. Stone, P. Albert, Llama 2: Open foundation and fine-tuned chat models, ArXiv abs/2307.09288 (2023). URL: https://api.semanticscholar.org/CorpusID:259950998.

[30] Mistral, Mistral large, our new flagship model, URL https://mistral.ai/news/mistral-large/, 2024.

[31] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, Mistral 7b, ArXiv abs/2310.06825 (2023). URL: https://api.semanticscholar.org/CorpusID:263830494.

[32] P. Cardinal-Fernandez, E. Garcia-Cuesta, J. Barberan, J. F. Varona, A. Estirado, A. Moreno, J. Villanueva, M. Villareal, O. Baez-Pravia, J. Menendez, et al., Clinical characteristics and outcomes of 1,331 patients with covid-19: Hm spanish cohort, Revista Española de Quimioterapia 34 (2021) 342.

[33] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text (2016) 2383–2392. doi:10.18653/v1/D16-1264.

[34] D. Chandrasekaran, V. Mago, Evolution of semantic similarity—a survey, ACM Computing Surveys 54 (2021) 1–37. URL: http://dx.doi.org/10.1145/3440755. doi:10.1145/3440755.

[35] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[36] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[37] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

A. Appendix. Detailed results of the consistency and veracity metrics for the three contexts.

Table A1
Consistency evaluation per question with *Context 1*

	GPT				Mistral				Gemini				Llama2			
	Semantic similarity	Overlap	ROUGE	BLEU	Semantic similarity	Overlap	ROUGE	BLEU	Semantic similarity	Overlap	ROUGE	BLEU	Semantic similarity	Overlap	ROUGE	BLEU
Q1	0,949	0,760	0,666	0,462	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q2	0,989	0,941	0,935	0,880	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,983	0,983	0,977
Q3	0,975	0,868	0,789	0,731	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q4	0,984	0,823	0,757	0,653	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,992	0,990	0,985
Q5	0,981	0,895	0,841	0,738	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q6	0,996	0,972	0,948	0,925	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q7	0,980	0,864	0,833	0,800	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q8	0,993	0,899	0,874	0,797	1,000	1,000	1,000	1,000	0,998	0,964	0,961	0,922	1,000	1,000	1,000	1,000
Q9	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q10	0,987	0,858	0,800	0,710	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Average	0,983	0,888	0,844	0,770	1,000	1,000	1,000	1,000	1,000	0,996	0,996	0,992	1,000	0,998	0,997	0,996

Table A2
Consistency evaluation per question with *Context 2*

	GPT				Mistral				Gemini				Llama2			
	Semantic similarity	Overlap	ROUGE	BLEU	Semantic similarity	Overlap	ROUGE	BLEU	Semantic similarity	Overlap	ROUGE	BLEU	Semantic similarity	Overlap	ROUGE	BLEU
Q1	0,976	0,969	0,924	0,825	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q2	0,997	0,990	0,988	0,976	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q3	0,972	0,869	0,832	0,663	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q4	0,998	0,926	0,920	0,875	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q5	0,997	1,000	0,963	0,904	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q6	0,939	0,706	0,556	0,383	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,991	0,969	0,957	0,925
Q7	0,949	0,736	0,649	0,475	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q8	0,996	0,932	0,896	0,867	0,991	0,954	0,927	0,903	0,995	0,960	0,957	0,925	0,995	0,944	0,932	0,903
Q9	0,990	0,927	0,906	0,862	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
Q10	0,982	0,918	0,895	0,796	1,000	0,991	0,989	0,977	0,998	1,000	0,975	0,938	1,000	1,000	1,000	1,000
Average	0,980	0,897	0,853	0,763	0,999	0,995	0,992	0,988	0,999	0,996	0,993	0,986	0,999	0,991	0,989	0,983

Table A3
Veracity evaluation per question with *Context 3*

	GPT				Mistral				Gemini				Llama2			
	Semantic similarity	Overlap	ROUGE	BLEU	Semantic similarity	Overlap	ROUGE	BLEU	Semantic similarity	Overlap	ROUGE	BLEU	Semantic similarity	Overlap	ROUGE	BLEU
Q1	0,777	0,750	0,466	0,612	0,770	0,750	0,424	0,537	0,412	0,250	0,181	0,368	0,777	0,750	0,466	0,612
Q2	0,647	0,554	0,331	0,347	0,677	0,550	0,244	0,298	0,596	0,599	0,461	0,496	0,668	0,700	0,363	0,294
Q3	0,687	0,285	0,178	0,289	0,681	0,380	0,173	0,349	0,698	0,500	0,266	0,358	0,680	0,380	0,210	0,420
Q4	0,764	0,448	0,207	0,409	0,805	0,466	0,239	0,507	0,714	0,157	0,093	0,218	0,648	0,290	0,163	0,304
Q5	0,665	0,428	0,255	0,184	0,619	0,312	0,222	0,259	0,622	0,350	0,260	0,137	0,634	0,297	0,101	0,271
Q6	0,778	0,458	0,368	0,479	0,737	0,333	0,170	0,384	0,774	0,533	0,294	0,242	0,787	0,500	0,307	0,436
Q7	0,758	0,350	0,191	0,373	0,733	0,256	0,161	0,361	0,631	0,413	0,254	0,306	0,821	0,435	0,242	0,454
Q8	0,889	0,679	0,543	0,769	0,791	0,411	0,235	0,440	0,858	0,529	0,461	0,673	0,926	0,647	0,156	0,405
Q9	0,705	0,269	0,202	0,196	0,725	0,242	0,107	0,259	0,746	0,230	0,163	0,220	0,691	0,222	0,175	0,300
Q10	0,731	0,420	0,264	0,423	0,728	0,333	0,244	0,402	0,708	0,588	0,299	0,266	0,706	0,441	0,210	0,437
Average	0,740	0,464	0,301	0,408	0,727	0,403	0,222	0,380	0,676	0,415	0,273	0,328	0,734	0,466	0,239	0,393