



Uncertain about ChatGPT: enabling the uncertainty evaluation of large language models

Anne-Laure Jousselme, Johan Pieter de Villiers, A. de Freitas, E. Blasch, Valentina Dragos, Gregor Pavlin, P. Costa, Kathryn B Laskey, C. Laudy

► To cite this version:

Anne-Laure Jousselme, Johan Pieter de Villiers, A. de Freitas, E. Blasch, Valentina Dragos, et al.. Uncertain about ChatGPT: enabling the uncertainty evaluation of large language models. 2023 26th International Conference on Information Fusion (FUSION), Jun 2023, Charleston, United States. <10.23919/FUSION52260.2023.10224086>. <hal-04627696>

HAL Id: hal-04627696

<https://hal.science/hal-04627696v1>

Submitted on 27 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Uncertain about ChatGPT: enabling the uncertainty evaluation of large language models

A-L. Jousselme¹, J.P. de Villiers², A. de Freitas², E. Blasch³, V. Dragos⁴, G. Pavlin⁵, P. C. Costa⁶,
K. B. Laskey⁶, C. Laudy⁷

¹ CS Group, La Garde, France, anne-laure.jousselme@csgroup.eu

²University of Pretoria, Pretoria, South Africa, pieter.devilliers@up.ac.za, allan.defreitas@up.ac.za

³Air Force Research Lab, Rome, NY, USA, erik.blasch.1@us.af.mil

⁴ONERA-The French Aerospace Lab, Palaiseau, France, valentina.dragos@onera.fr

⁵D-CIS Lab, Thales Research and Technology, Delft, The Netherlands, gregor.pavlin@d-cis.nl

⁶George Mason University, Fairfax, VA, USA, [pcosta,klaskey]@gmu.edu

⁷Thales, Palaiseau, France, claire.laudy@thalesgroup.com

Abstract—ChatGPT, OpenAI’s chatbot, has gained considerable attention since its launch in November 2022, owing to its ability to formulate articulated responses to text queries and comments relating to seemingly any conceivable subject. As impressive as the majority of interactions with ChatGPT are, this large language model has a number of acknowledged shortcomings, which in several cases, may be directly related to how ChatGPT handles uncertainty. The objective of this paper is to pave the way to formal analysis of ChatGPT uncertainty handling. To this end, the ability of the Uncertainty Representation and Reasoning Framework (URREF) ontology is assessed, to support such analysis. Elements of structured experiments for reproducible results are identified. The dataset built varies Information Criteria of Correctness, Non-specificity, Self-confidence, Relevance and Inconsistency, and the Source Criteria of Reliability, Competency and Type. ChatGPT’s answers are analyzed along Information Criteria of Correctness, Non-specificity and Self-confidence. Both generic and singular information are sequentially provided. The outcome of this preliminary study is twofold: Firstly, we validate that the experimental setup is efficient in capturing aspects of ChatGPT uncertainty handling. Secondly, we identify possible modifications to the URREF ontology that will be discussed and eventually implemented in URREF ontology Version 4.0 under development.

Index Terms—Uncertainty evaluation; Ontology; Information quality; Source quality; Large Language Models; NLP.

I. INTRODUCTION

ChatGPT [1] is a chatbot developed by the company OpenAI, and was launched on 30 November 2022. It is based on the generative pre-trained transformer (GPT) model (version 3.5), which is a sophisticated large-scale machine learning language model. ChatGPT has gained considerable attention since its launch, owing to its ability to formulate articulate responses to text queries and comments relating to seemingly any conceivable subject [2]. The language model provides credible answers and early adopters consider ChatGPT as *the most disruptive technology* [3] of late. As impressive as the majority of interactions with ChatGPT are, even OpenAI admit that the model has some shortcomings [1]. Notably, ChatGPT i) sometimes writes plausible-sounding but incorrect

or nonsensical statements, ii) is sensitive to perturbations in the input phrasing which affect the chatbot’s ability to formulate a response, iii) is often excessively verbose and overuses certain phrases, iv) tends to jump to an interpretation of the intended query, rather than asking for clarity, and v) can sometimes respond to inappropriate questions or exhibit bias (*e.g.*, [4]).

These shortcomings may in several cases be directly related to how ChatGPT handles uncertainties in its input, output and inference. These uncertainties are also a direct result of how the underlying model was trained, and as such, uncertainties that enter during the modeling process should also be considered (see for example [5], [6] where modeling considerations are discussed). The Uncertainty Representation and Reasoning Framework (URREF) [7], [8] enables the evaluation of uncertainty in the inputs, reasoning and outputs of fusion systems, and is in continuous development by the official International Society of Information Fusion (ISIF) Evaluation of Techniques for Uncertainty Representation and Reasoning Working Group (ETURWG) since 2011. We assert here that concepts and criteria from the URREF ontology can be leveraged to support the uncertainty evaluation of large language models (LLMs) when asked to perform an information fusion task.

The objective of this paper is thus twofold: (1) to show how the URREF ontology can support first order empirical attempt at evaluating uncertainty representation and reasoning of ChatGPT, while prospective classes of the ontology are used, and (2) to propose further discussions for the ETURWG around the enrichment or adaptation of the URREF ontology. The goal of uncertainty analysis is to go beyond making an inventory of errors [9]. Thus, by using a well-established uncertainty framework, it becomes possible to gain a deeper understanding of the various types of uncertainty affecting the LLM answers. The URREF will be used as the reference uncertainty framework to capture and describe the nature of uncertainty affecting the ChatGPT-human dialogue. Hence, rather than characterizing ChatGPT’s internal uncertainty representation, we will assess the impact of imperfect information

on its uncertain answers.

The paper is structured as follows: In Section II, we explore some functionalities of ChatGPT and review some Natural Language Processing (NLP) techniques in the fusion community. We propose in Section III a structured methodology to assess uncertainly handling. Section IV reports some results of the experiments and we conclude in Section V on future work.

II. BACKGROUND

A. NLP approaches in information fusion

For the last 15 years at least, NLP approaches have been part of the information fusion paradigm. Since 2008, Data Fusion for NLP has been widely explored as combining knowledge from different documents supporting both the quality and quantity of textual data (*i.e.*, soft data from human reports). Furthermore, NLP has also been used for information fusion terminology analysis, query expansion, and document retrieval [10] and to provide context for situational analysis of sensor results (*e.g.*, Automatic Target Recognition) for human discourse [11]. Auger and Roy [12] looked at uncertainty from linguistic and reference ambiguity, seeking to develop a probabilistic numerical tagging based on semantic analysis. Standard knowledge-base design with rules of data combination develop methods for syntactic and semantic analysis for soft (text) data fusion such as context-based information retrieval (CBIR) [13].

With advances in machine learning (ML), fusion of techniques for NLP were investigated such as a Bayes' classifier for sentiment analysis [14], and word search [15]. Sun *et al.* [16] highlight the many developments for NLP ML including graphical Latent Dirichlet Allocation (LDA), machine learning (*e.g.*, Support Vector Machine), opinion mining and sentiment analysis by extracting features, including tokenization, word segmentation, Part of Speech (POS) tagging, and parsing. Among the many supervised and unsupervised approaches, data fusion was highlighted in domain adaptation and probabilistic generative model-based approaches.

Deep learning (DL) offers much promise for multimodal methods that incorporates NLP data fusion analysis such as the popular Bidirectional Encoder Representations from Transformers (BERT) approach [17] that motivated the interest in tools like ChatGPT. You *et al.* [18] highlight DL methods for NLP using data fusion where a transformer layer on top of BERT layer captures document features and generates the corresponding summary. A review of multimodal intelligent systems summarizes methods for operation-based fusion, attention-based fusion, and bilinear pooling-based fusion for NLP and question-answering methods that include multimodal data (*e.g.*, video and text) [19].

Physical and semantic sensor data fusion (SDF) can be enhanced with AI/ML/DL systems such as for analysis of twitter data [20], uncertainty analysis of social data [21] and word convolution-deconvolution vector fusion word embedding for polysemous and task-aware NLP [22]. Other examples highlight use of models, human-aware ML, evidential neural

networks, and edge devices for intelligibility, interpretability and explainability of results [23]. Consistently, NLP review papers highlight input and output fusion techniques using entity feature embedding and contextual embedding, respectively for entity, relationship, and narrative content summarization [24].

B. Exploring ChatGPT functionalities

Despite the recent release of ChatGPT, the research community already reports several initiatives investigating various aspects of ChatGPT. Research studies fall under the following categories: comparing ChatGPT answers and human expertise, comparing results of ChatGPT and other language models, and analyzing the effectiveness of ChatGPT in specific tasks.

One of the first experiments was designed to analyze how close ChatGPT is to human experts, and more specifically how helpful, honest or harmless the model is when compared to human experts [25]. Although ChatGPT's answers were considered to be generally more helpful than those of human, it struggled with some specialized domains and was shown to produce harmful and biased responses in some cases.

In some applications such a plagiarism detection, it may be important to discriminate between human and LLM generated text. In [26] a transformer-based model was used to discriminate between human-written and ChatGPT-generated short text. The analysis of the classification results (with an accuracy of 79%) highlighted several distinctive features of ChatGPT answers: the absence of specific details, the use of atypical, impersonal vocabulary, and the lack of feelings.

The authors of [27] evaluate the capabilities of ChatGPT in even NLP tasks, including question answering, summarization, named entity recognition, dialogue, natural language inference, basic reasoning and sentiment analysis. The results indicate that ChatGPT is a powerful generalist model, having good reasoning and dialogue capabilities, while still facing challenges when solving specific tasks, such as summarizing texts.

The capability of ChatGPT to formulate Boolean queries for systematic literature reviews was analyzed by Wang and colleagues in [28] based on information retrieval results from PubMed's Entrez, using an API to retrieve PubMed IDs [29]. Results of evaluation show that ChatGPT obtained higher precision but lower recall values than queries generated by state-of-the-art automatic query formulation methods. An identified shortcoming is that ChatGPT generates different queries even if the same prompt is used, which may adversely affect the effectiveness of query generation.

The mathematical abilities of ChatGPT have been investigated in [30], within the context of answering exam-style mathematical questions in a number of mathematical contexts. The results indicate that ChatGPT achieves poor performance for most difficult mathematical problems. Moreover, although the model offers insightful proofs in a few cases, ChatGPT is currently unable to deliver high-quality proofs or calculations consistently. Since ChatGPT is a language model, and according to the "no free lunch" assertion, it cannot be expected to outperform models that are task specific, such as competency in mathematics.

The above studies emphasize that as it currently stands, human validation of ChatGPT results and the mental cost of this validation can, in some applications, outweigh the advantages that the model offers [31]. Yet, no study so far focused on the fusion abilities of ChatGPT and its uncertainty handling.

C. ChatGPT reasoning capabilities

As a LLM, ChatGPT aims at predicting the statistical most probable follow-up of bits of texts, or answers to questions, according to its training data. Although ChatGPT may seem to have reasoning capabilities, it is not equipped with logical reasoning capabilities as are rules- or logic-based reasoning engines. ChatGPT, as other LLMs such as [32] may be trained to answer logical questions, through the adaptation of the training data to learn on specific tasks. Therefore, the model does not perform inferences by itself, but rather mimics what it encountered during the training phase. This feature may be useful in situation when decision support system developers do not have access to a logical rule base. When assessed on both popular and distribution benchmarks, ChatGPT's logical abilities have been shown to be higher on old popular benchmarks than on recent ones while improved when context is provided (*e.g.*, [33], [34]). In this paper, we propose to evaluate how chatGPT deals with uncertainty in fusion tasks, that can be seen as contextual reasoning questions. We do not imply that chatGPT has the capability to handle fusion tasks as precisely as a well-defined and adapted fusion system would do. However, LLMs are becoming an integral part of our lives, we want to evaluate how they could be useful to the fusion scientific community, for example as an additional NLP component of a fusion system.

Capabilities of ChatGPT were mostly considered from domain specific perspectives and with limited evaluation of the uncertainty representation and reasoning within ChatGPT. In the next section, we propose a structured methodology which enables assessing ChatGPT's abilities in handling uncertainty. We essentially pave the way for future larger evaluations. We use the URREF criteria as defined in [8], with additional proposals that aim at covering the wide variety of sources and processing methods in information fusion as far as uncertainty handling is concerned.

III. STRUCTURED METHODOLOGY TO UNCERTAINTY HANDLING EVALUATION

Rather than an exhaustive uncertainty evaluation, within the constraints of this paper the basic elements of a general methodology are established to be further exploited. Furthermore, the ability of the URREF ontology to support experiments and to evaluate how ChatGPT handles uncertainty, is determined. To this end, we adapted the experiment originally designed for the Risk Game [35], to provide ChatGPT with a structured set of pieces of information of varying quality, and we analyzed the answers. The example evaluation is performed by varying a small subset of INFORMATION CRITERIA and dual UNCERTAINTY TYPES as defined within the URREF

ontology [8]. Referring to the classes presented below, the class INFORMATION is a new element to the URREF, the class SOURCE CRITERIA has been populated with instances and the class INFORMATION CRITERIA is an update of the current class.

The purpose of the Risk Game was to study the impact of information quality and uncertainty on human belief assessment and decision making [35]. Players were asked to find a missing vessel among two possible options (Vessel *A* or Vessel *B*), based on a series of pieces of information of variable quality queried through various sources (sensors, AI agents and humans). We will reuse and adapt here the structure of this dataset to stress ChatGPT on its ability to handle imperfect information, and perform basic fusion tasks. Equivalently to information provided to players of the Risk Game, we will provide ChatGPT with varying pieces of information and capture its answers.

A. The maritime surveillance scenario

The scenario is the following [35]: The Automatic Identification System (AIS) contact with a particular fishing vessel was lost one hour ago, leaving the Watch Officer (WO) from *LeftLand* with two unidentified vessels (or tracks), say Vessel *A* and Vessel *B*. The question the WO has to answer is "Which of the two vessels, Vessel *A* or Vessel *B*, is the missing vessel?". The scenario is sketched in Figure 1.

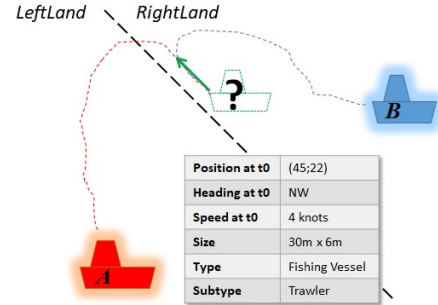


Fig. 1: The last AIS information sent by the Missing Vessel (in the gray table) and possible current locations identified as Vessel *A* in LeftLand or Vessel *B* in RightLand

B. Sources

Sources available to the WO combine a variety of sensors both cooperative (*e.g.*, AIS) and non-cooperative (*e.g.*, radar, camera), whose measurement is processed either by algorithms (*e.g.*, tracker, Automatic Target Recognition (ATR) algorithm) or humans (*e.g.*, camera analyst, cargo captain). The radar covers the whole area, the Infra-Red (IR) camera covers only the area around Vessel *A*, a cargo captain is in the vicinity of Vessel *B* but too far from Vessel *A* for visual identification, and a Synthetic Aperture Radar (SAR) imagery covering the whole area has been taken 30 minutes ago.

To characterize the sources of information, we refer to the list of criteria in Fig. 3a. The VERACITY is a measure of the extent to which a source reports what it assesses to be

the case [8]. The SELF-CONFIDENCE is a measure of the source's own credibility [8], *i.e.*, how much it believes (or is confident) in its statement. The COMPETENCY of a source is essentially characterized by its coverage, *i.e.* whether it sees Vessel A, Vessel B or both, and to the attribute it reports about (*i.e.*, location, speed, heading, size or type). For instance, the radar operator is competent for providing information about the two vessels (wide coverage) and information about the size of the vessel. But he/she is not competent to provide information about the speed. The REPUTATION is understood here as related essentially to the type of source (*i.e.*, sensor, human, combination of both) and not explicitly mentioned. RELIABILITY expresses how we should consider (rely on) the information provided by the source. It is usually assessed based on past experience with the source, where it has been verified the validity (truth) of the statements provided. Although some links may exist between these criteria (*e.g.*, the VERACITY is impacted by the COMPETENCY of the source, RELIABILITY is impacted by VERACITY and COMPETENCY), they will not be considered in the following.

C. Information

Information is split into two categories of KNOWLEDGE (or generic information) and EVIDENCE (or singular information) following [36], while a possible typology is provided in Fig. 2. Essentially, generic information applies to a population,

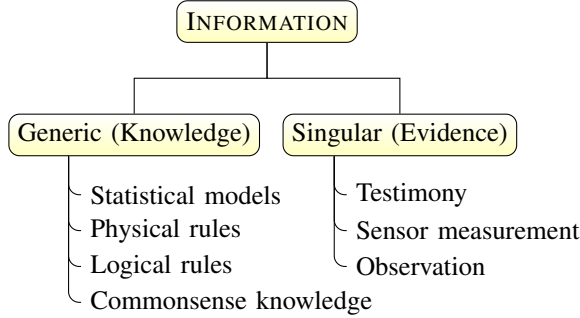


Fig. 2: Class INFORMATION and some sub-categories, as inspired by [36]

while singular information applies to an individual.

Prior information: Prior information include on the one hand, singular information about the missing vessel as provided by its last AIS contact: its name, Maritime Mobile Service Identity (MMSI), International Maritime Organization (IMO), type, length, width, etc, are known with a very high confidence to the WO. On the other hand, some generic knowledge about the traffic in the area, about the reliability and competency of the sources, some logical rules, and information about the geographical area (*e.g.*, some distance to sensors).

Table I summarizes prior information as a series of sets of pieces of information, named $\Phi_s^{(i)}$ for singular information and $\Phi_g^{(j)}$ for generic information. These sets of information can be varied for the different experiments.

TABLE I: Generic and singular prior information

$\Phi_s^{(1)}$ SINGULAR INFORMATION - SET UP
The missing vessel is either Vessel A or Vessel B
$\Phi_s^{(2)}$ SINGULAR INFORMATION - PRIOR ABOUT MISSING VESSEL
The missing vessel is a fishing vessel
The missing vessel is a trawler
The length of the missing vessel is 30 meters
The width of the missing vessel is 6 meters
$\Phi_g^{(1)}$ GENERIC INFORMATION - LOGICAL RULES
If Vessel A is not a fishing vessel then the missing vessel is Vessel B
If Vessel B is not a fishing vessel then the missing vessel is Vessel A
$\Phi_g^{(2)}$ GENERIC INFORMATION - SOURCE COMPETENCY (COVERAGE)
A radar with a tracker reports target (vessel) locations
A radar operator Ben reports on target heading and size
A radar automatic target recognition system reports on target type
A SAR sensor operator Chris reports on target location and heading
A SAR automatic target recognition system reports on target size
An IR camera analyst Anna reports on target location, speed, heading, type and size
A cargo vessel captain Daniel reports on target location, speed, heading, type and size

Current information: Sources provide imperfect information about the two vessels. Three quality dimensions from the information criteria listed in Fig. 3b will be made explicitly varying at the input, namely CORRECTNESS, UNCERTAINTY, NON-SPECIFICITY. Hence, information provided can be either

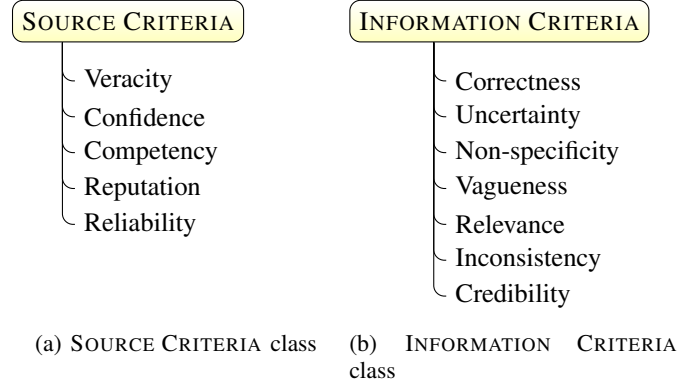


Fig. 3: Criteria classes used in the maritime use case evaluation of ChatGPT

true or false, certain or uncertain (in this case, as only related to the source SELF-CONFIDENCE about its estimation), and more or less specific (*i.e.*, bearing on several possible values for the attribute estimated). Human sources can provide information in natural language, and thus VAGUENESS will be considered as a special kind of NON-SPECIFICITY. Additional latent criteria exist: Because of the lack of correctness, combination of information can be INCONSISTENT (or conflicting). Besides, some information is not RELEVANT to solve the problem at hand. For instance, the location and the heading do not help to match the two tracks to the identity of the missing vessel. Finally, CREDIBILITY is a higher-level criterion referring to the degree to which a piece of information can be believed or

accepted as true, real, or honest [8].

D. ChatGPT instruction

Table II presents the instruction provided to ChatGPT. For the following experiments, we will “force” ChatGPT

TABLE II: ChatGPT Instruction

Φ_I INSTRUCTION
Given the information provided, you should choose one statement as your most confident response. The three statements are:
1) “the missing vessel is Vessel A”,
2) “the missing vessel is Vessel B”,
3) “the missing vessel is Vessel A or Vessel B”.
Firstly, respond with one of the above statements in which you have the highest confidence, preceded by nothing else. Secondly, you should then respond with a confidence level for each of the three statements 1) to 3). Respond with a confidence level on a scale of 1 to 6 where 1 represents very low confidence, 2 represents low confidence, 3 represents medium low confidence, 4 represents medium high confidence, 5 represents high confidence, and 6 represents very high confidence. Please respond with the following format: “<confidence that Vessel A is the missing vessel = x>”, “<confidence that Vessel B is the missing vessel = y>”, and “<confidence that Vessel A or Vessel B is the missing vessel = z>” where x is the confidence of the missing vessel being Vessel A, y is the confidence of the missing vessel being Vessel B and z is the confidence of the missing vessel being Vessel A or Vessel B. Thirdly, The response should be followed by an explanation of the reasoning behind the answer.

to express its confidence in the answer provided on a Likert scale between 1 and 6, where 6 means maximal confidence and 1 minimal confidence. Besides, we will allow ChatGPT to provide some confidence toward hypotheses *A* only, *B* only and *A* or *B*. This last choice is here to provide some freedom in expressing uncertainty, compliant with non-additive uncertainty theories. These instructions are seen as another set of items Φ_I that can be made varying.

IV. EXPERIMENTS AND RESULTS

We conducted a series of experiments which evaluated how uncertainty is handled from the following responses based on:

- 1) sets of prior information, generic and singular;
- 2) unique target attributes based information items of varying quality.

In these experiments, we evaluate information generated by ChatGPT as a response to a prompt. Evidence about Vessels *A* and/or *B* are appended to the prior information from Table I, followed by the instruction of Table II. ChatGPT is prompted to provide a numerical degree of belief on a Likert scale with 6 levels between 1 and 6 as described in Table II. The pieces of evidence are varied along information quality dimensions from Fig. 3b. These criteria apply to unique information items, namely CORRECTNESS, *i.e.*, whether the source correctly identifies Vessel *A* as the missing vessel, UNCERTAINTY, *i.e.*, how much the source is certain of its answer and NON-SPECIFICITY, *i.e.*, if it provides a single value for an attribute as an answer or whether it provides an interval of values. It is important to notice that the UNCERTAINTY degree output by ChatGPT corresponds to the SELF-CONFIDENCE criterion of Fig. 3a, when ChatGPT would be considered as a source of information.

A. Background knowledge experiments

In the first experiment, we successively provide ChatGPT with the following sets of prior information (ref. Table I):

- 1) $\Phi_1 = \{\Phi_s^{(1)}\}$,
- 2) $\Phi_2 = \{\Phi_s^{(1)}, \Phi_s^{(2)}\}$,
- 3) $\Phi_3 = \{\Phi_s^{(1)}, \Phi_s^{(2)}, \Phi_g^{(1)}\}$,
- 4) $\Phi_4 = \{\Phi_s^{(1)}, \Phi_s^{(2)}, \Phi_g^{(1)}, \Phi_g^{(2)}\}$,

followed by the ChatGPT instruction in Table II.

TABLE III: Excerpt of ChatGPT responses for varying prior information. Interesting Remarks from ChatGPT are highlighted in bold.

	EXCERPT OF CHATGPT RESPONSES
Φ_1	“...I have assigned a confidence level of 3 to each statement “the missing vessel is Vessel A” and “the missing vessel is Vessel B” because although it is possible that the missing vessel is either of them, there is not enough information to definitively conclude that one is more likely than the other. ”
Φ_2	“...Without that additional information, I can only be moderately confident that the missing vessel is either Vessel A or Vessel B. ”
Φ_3	“... Since we cannot eliminate either vessel as a possibility , our most confident response is that the missing vessel is either Vessel A or Vessel B.”
Φ_4	“...My confidence level that the missing vessel is Vessel A is low (1) because we know that Vessel A is not a fishing vessel , and the missing vessel is a fishing trawler. Additionally, there is no other information that suggests that Vessel A could be the missing vessel.”

Since no evidence has been provided to support any of the two hypotheses *A* or *B*, it is expected that ChatGPT would respond with an equal confidence in the statements “the missing vessel is Vessel *A*” and “the missing vessel is Vessel *B*” and a larger confidence in “the missing vessel is either Vessel *A* or Vessel *B*”. The results indicate that this was indeed the case for Φ_1 , Φ_2 and Φ_3 . In these cases, ChatGPT’s explanation also coincides with the logical expectation as indicated in Table III. However, in the case of Φ_4 , ChatGPT erroneously indicates the highest confidence in the statement “the missing vessel is Vessel *B*”. Analysing the response, as indicated in the fourth item of Table III, it was found that ChatGPT has erroneously extracted prior information from $\Phi_g^{(1)}$ as evidence, specifically focusing on the text in bold in the statement: “If **Vessel A is not a fishing vessel** then the missing vessel is Vessel *B*”.

To reduce the likelihood of erroneous evidence extraction, prior information was restricted to Φ_2 and a subset of $\Phi_g^{(2)}$ related to the evidence provided in the experiments following this section. This prior information was confirmed to produce the logically expected response from ChatGPT.

B. Information fusion experiments - Varying quality

In this experiment, a ChatGPT prompt consists of two pieces of information, one relating to Vessel *A* and the other relating to Vessel *B* along with the prior information of Table I and the instruction of Table II. The results of several such prompts are contained in Fig. 4 to Fig. 6. The information quality is

varied along the dimensions CORRECTNESS, UNCERTAINTY and NON-SPECIFICITY. In Fig. 4 to Fig. 6, the information quality is depicted on the y-axis, where, “Spe” denotes specific and relates to NON-SPECIFICITY, “Cer” denotes certain and relates to UNCERTAINTY, and “¬” denotes a logical NOT. The top half of the plot corresponds to CORRECT evidence and the bottom half to INCORRECT evidence. ChatGPT’s confidence in Vessel A is displayed in red, in Vessel B in blue and in either Vessel A or Vessel B in green. Different attributes are provided, more or less informative (e.g., discriminative). If incorrect evidence is provided to ChatGPT, then it may happen that it reaches incorrect conclusion, except when incorrect evidence supports the correct conclusion. As an interesting side note, upon asking ChatGPT what type of reasoning it used to arrive at a response, it responded that it uses abductive reasoning. However, this may be taken with a grain of salt, since we lack information about how ChatGPT handles its own limitations.

Fig. 4 depicts the reasoning results if a discriminative attribute is used for providing evidence¹. The fact that size based evidence for both vessels is provided, makes it easier for ChatGPT to match each of Vessels A or B to the dimensions of the missing vessel. In all cases where incorrect information is provided, ChatGPT arrives at the incorrect conclusion, as would be generally expected. In cases where correct information is provided, ChatGPT arrives at the correct conclusion in three out of four cases. The exception is when non-specific evidence is provided to ChatGPT (third row of Fig. 4). Non-specific evidence implies an interval of vessel dimensions. In certain cases it was observed that non-specific evidence negates the discriminative quality of the information provided, as the vessel sizes are not that far apart (e.g., “between 35 and 45 meters”, while the real length of Vessel A is 30 meters). However, in this specific instance, it was noted that ChatGPT incorrectly excluded the width of the missing vessel from the range provided for the width of vessel A while simultaneously including the width of the missing vessel in the range provided for the width of vessel B as indicated in Table IV. It is interesting to note that this behaviour did not occur in the case of incorrect information, where ChatGPT’s responses were in-line with expectation.

Fig. 5 depicts the reasoning results if the **type** attribute is used for providing evidence about the two vessels. For example, evidence for both Vessels A and B being fishing vessels are provided as indicated in Table V, or evidence for both Vessels A and B being recreational vessels are provided in the case of incorrect information. This means that there is no discriminative evidence, irrespective of whether it is correct or incorrect. As such, ChatGPT takes the safest option and provides the highest confidence in the statement “the missing vessel is Vessel A or Vessel B”.

Fig. 6 depicts the reasoning results if a poorly discriminative attribute, without any link to the provided prior information,

¹In this scenario, discriminative attributes are essentially the size (width and length) as both vessel A and B are fishing vessels, trawlers, Vessel B being smaller than Vessel A.

TABLE IV: Excerpt of ChatGPT response for the case of correct, non-specific and certain evidence provided for Vessel A and B for the size attribute. This corresponds to the third line of results in Figure 4. Illogical remarks from ChatGPT are highlighted in bold.

EVIDENCE PROVIDED TO CHATGPT

- According to the SAR automatic target recognition algorithm, the length of Vessel A is between 25 and 40 meters.
- According to the SAR automatic target recognition algorithm, the length of Vessel B is between 15 and 30 meters.
- According to the analyst Anna, the width of Vessel A is between 5 and 8 meters.
- According to the cargo captain Daniel, the width of Vessel B is between 3 and 5 meters.

EXCERPT OF CHATGPT RESPONSE

My confidence level for Vessel A being the missing vessel is low (2) because although the length range provided by the SAR automatic target recognition algorithm includes the missing vessel’s length, **the width range reported by the analyst Anna is not a good match. The width of the missing vessel is reported to be 6 meters, which falls outside the width range of Vessel A (5-8 meters).** Therefore, it is unlikely that Vessel A is the missing vessel.

My confidence level for Vessel B being the missing vessel is high (5) because **the width range reported by the cargo captain Daniel matches the width of the missing vessel**, and the length range provided by the SAR automatic target recognition algorithm also includes the missing vessel’s length. The reported length range of Vessel B (15-30 meters) includes the length of the missing vessel, and the reported width range of Vessel B (3-5 meters) matches the width of the missing vessel. Therefore, it is likely that Vessel B is the missing vessel.

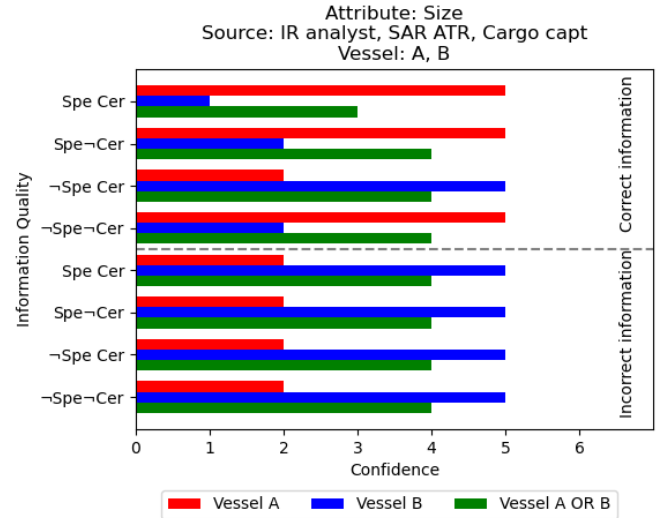


Fig. 4: Fusion experiments - Informative attribute: The sizes of both vessel A and vessel B are reported.

(in this case **speed**) is used for providing evidence. It is noted that in the case of both specific and certain evidence, the response is in-line with expectation. However, fluctuations away from expectation begin to occur once non-specific and uncertain evidence is provided. In all of these instances, ChatGPT’s motivation for a greater confidence in either vessel

TABLE V: Excerpt of ChatGPT response for the case of correct, specific and certain evidence provided for Vessel A and B for the type attribute. This corresponds to the first line of results in Figure 5.

EVIDENCE PROVIDED TO CHATGPT

- According to the radar automatic target recognition algorithm, the type of Vessel A is a Fishing vessel.
- According to the radar automatic target recognition algorithm, the type of Vessel B is a Fishing vessel.

EXCERPT OF CHATGPT RESPONSE

“...There are a few pieces of information provided that can help us narrow down the possibilities for the missing vessel. We know that the missing vessel is a fishing vessel of subtype trawler, and it has a length of 30 meters and a width of 6 meters. Both Vessel A and Vessel B are identified by the radar automatic target recognition algorithm as fishing vessels, which is consistent with the subtype trawler. However, the algorithm does not provide any additional information about the vessels, such as their size or location.”

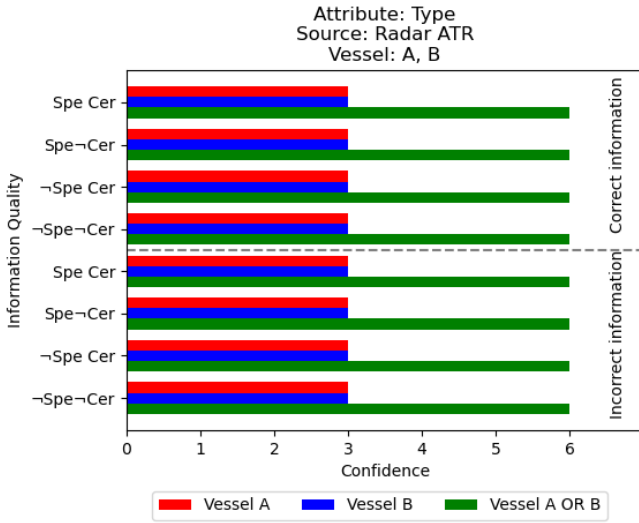


Fig. 5: Fusion experiments - Same vessel type reported for A and B: Vessel A is always reported as the same type as Vessel B, either correctly or incorrectly.

is based on ChatGPT relying on prior knowledge it gained through model training. In this case, the average speed of specific fishing vessels as illustrated in Table VI.

V. CONCLUSIONS

The ChatGPT LLM is currently widely considered to be a disruptive technology, and its potential impacts are largely discussed in mainstream media. Since LLMs are taking center stage in the AI community, the URREF ontology design to support the assessment of uncertainty handling in fusion systems, should be able to also support the evaluation of uncertainties in the inputs, representation and reasoning and outputs of LLMs. In this paper, we proposed a methodology leveraging from the URREF ontology, identifying the elements for structured and reproducible experiments to assess

TABLE VI: Excerpt of ChatGPT response for the case of correct, non-specific and certain evidence provided for Vessel A and B for the speed attribute. This corresponds to the third line of results in Figure 6. Remarks from ChatGPT indicating prior information from outside of the experiment are highlighted in bold.

EVIDENCE PROVIDED TO CHATGPT

- According to the radar tracker, the speed of Vessel A is between 5 and 7 knots.
- According to the radar tracker, the speed of Vessel B is between 6 and 8 knots.

EXCERPT OF CHATGPT RESPONSE

“...I am more confident that the missing vessel is Vessel B based on the radar tracker’s report that its speed is between 6 and 8 knots, **which is a closer range to the typical speed of a fishing trawler vessel. The radar tracker’s report on Vessel A’s speed is also within a reasonable range, but the reported speed range of Vessel B is a better fit for a fishing trawler vessel.** However, there is still a chance that the missing vessel could be either Vessel A or Vessel B, since the information provided is limited and does not provide a definitive answer. Therefore, I have medium-high confidence that the missing vessel is either Vessel A or Vessel B.”

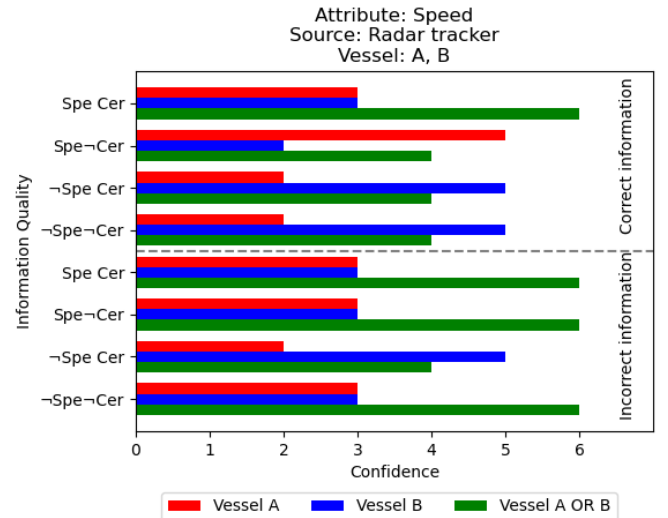


Fig. 6: Fusion experiments - Uninformative attribute: The speed of both vessels is reported.

uncertainty handling of ChatGPT. As an outcome, we propose further discussions for the ETURWG around the enrichment or adaptation of the URREF ontology in relation to its applicability to LLMs, by proposing or updating classes SOURCE CRITERIA, INFORMATION CRITERIA and INFORMATION (type). Thanks to the generality of the URREF ontology, the proposed methodology does not only enable structured and reproducible evaluation experiments of ChatGPT in its ability to handle uncertainty, but it can be transferred to the evaluation of other LLMs.

The methodology is illustrated on a maritime use case. ChatGPT is provided with generic contextual background

knowledge about a surveillance scenario, singular pieces of evidence that need to be fused, and an instruction requiring a structured answer on a Likert scale. Explanation leading to the answer is also requested. The results indicated that ChatGPT provides answers consistent with the input information, if the input information is correctly interpreted by ChatGPT. In addition, it is important to be cognisant of the possibility that responses from LLMs could potentially be impacted by the prior knowledge which the model has gained through model training. This prior knowledge is typically unknown to end-users of the LLMs and can lead to unexpected responses. This refers to the DATA HANDLING CRITERIA of the URREF. This paper is not only a tentative start on the journey towards a systematic approach for evaluating LLMs supported by the URREF ontology, but considers how the URREF ontology can be modified in light of such evaluation experiments.

Future work will include investigations into how LLMs such as ChatGPT handle irrelevant information, conflict between sources, as well as knowledge about source quality. Furthermore, evaluation over multiple Monte Carlo runs will be required to assess the determinism of ChatGPT, as captured by some URREF criteria such as CONSISTENCY.

REFERENCES

- [1] "ChatGPT Blog Website." [Online]. Available: <https://openai.com/blog/chatgpt/>
- [2] K. Roose, "The Brilliance and Weirdness of ChatGPT." [Online]. Available: <https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html>
- [3] M. U. Haque, I. Dharmadasa, Z. T. Sworna, R. N. Rajapakse, and H. Ahmad, "i think this is the most disruptive technology": Exploring sentiments of ChatGPT early adopters using twitter data," *arXiv preprint*, vol. 2212.05856, 2022.
- [4] K. Chen, A. Shao, J. Burapachee, and Y. Li, "A critical appraisal of equity in conversational ai: Evidence from auditing gpt-3's dialogues with different publics on climate change and black lives matter," *arXiv preprint*, vol. 2209.13627, 2022.
- [5] J. P. de Villiers, K. Laskey, A.-L. Jousselme, E. Blasch, A. de Waal, G. Pavlin, and P. Costa, "Uncertainty representation, quantification and evaluation for data and information fusion," in *Int. Conf. on Information Fusion*, Washington DC, USA, July 2015, pp. 50–57.
- [6] J. De Villiers, G. Pavlin, A.-L. Jousselme, S. Maskell, A. de Waal, K. Laskey, E. Blasch, and P. Costa, "Uncertainty representation and evaluation for modeling and decision-making in information fusion," *Journal for Advances in Information Fusion*, vol. 13, no. 2, pp. 198–215, 2018.
- [7] P. C. G. Costa, K. B. Laskey, E. Blasch, and A.-L. Jousselme, "Towards unbiased evaluation of uncertainty reasoning: The URREF ontology," in *Int. Conf. on Information Fusion*, Singapore, 2012, pp. 2301–2308.
- [8] P. Costa, A.-L. Jousselme, K. Laskey, E. Blasch, V. Dragos, J. Ziegler, P. de Villiers, and G. Pavlin, "Urref: Uncertainty representation and reasoning evaluation framework for information fusion," *Journal of Advances in Information Fusion*, vol. 13, no. 2, pp. 137–157, 2019.
- [9] A. Borji, "A categorical archive of chatgpt failures," *arXiv preprint arXiv:2302.03494*, 2023.
- [10] E. Dura and B. Gawronska, "Natural language processing in information fusion terminology management," in *International Conference on Information Fusion*, Cologne, GE, July 2008.
- [11] A. Steinberg and G. Rogova, "Situation and context in data fusion and natural language understanding," in *Int. Conf. on Information Fusion*, Cologne, GE, July 2008.
- [12] A. Auger and J. Roy, "Expression of uncertainty in linguistic data," in *Int. Conf. on Information Fusion*, Cologne, GE, July 2008.
- [13] S. Shapiro and D. Schlegel, "Natural language understanding for soft information fusion," in *Int. Conf. on Information Fusion*, Istanbul, TR, July 2013, pp. 380–388.
- [14] B. Liu, E. Blasch, Y. Chen, D. Shen, and G. Chen, "Scalable sentiment classification for big data analysis using naïve bayes classifier," in *IEEE International Conference on Big Data*, 2013, pp. 99–104.
- [15] S. Lakhanpal, A. Gupta, and R. Agrawal, "Discover trending domains using fusion of supervised machine learning with natural language processing," in *Int. Conf. on Information Fusion*, Washington DC, USA, July 2015, pp. 893–900.
- [16] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information Fusion*, vol. 36, pp. 10–25, 2017.
- [17] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.
- [18] F. You, S. Zhao, and J. Chen, "A topic information fusion and semantic relevance for text summarization," *IEEE Access*, vol. 8, pp. 178 946–178 953, 2020.
- [19] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, March 2020.
- [20] L. Evans, M. Owda, K. Crockett, and A. F. Vilas, "A methodology for the resolution of cashtag collisions on twitter - a natural language processing & data fusion approach," *Expert Syst. Appl.*, vol. 127, pp. 353–369, 2019. [Online]. Available: <https://doi.org/10.1016/j.eswa.2019.02.035>
- [21] C. Laudy and V. Dragos, "Use cases for social data analysis with URREF criteria," in *Int. Conf. on Information Fusion*, Virtual, July 2020.
- [22] K. Shuang, Z. Zhang, J. Loo, and S. Su, "Convolution-deconvolution word embedding: An end-to-end multi-prototype fusion embedding method for natural language processing," *Information Fusion*, vol. 53, pp. 112–122, 2020.
- [23] E. Blasch, T. Pham, C.-Y. Chi, and et al., "Machine learning/artificial intelligence for sensor data fusion—opportunities and challenges," *IEEE Aerospace and Electronic Systems Magazine*, vol. 36, pp. 80–93, 2021.
- [24] Y. Wang, W. Wang, Q. Chen, K. Huang, A. Nguyen, S. De, and A. Hussain, "Fusing external knowledge resources for natural language understanding techniques: A survey," *Information Fusion*, vol. 92, pp. 190–204, 2023.
- [25] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How close is ChatGPT to human experts? comparison corpus, evaluation, and detection," *arXiv preprint*, vol. 2301.07597, 2023.
- [26] S. Mitrović, D. Andreoletti, and O. Ayoub, "ChatGPT or human? detect and explain. explaining decisions of machine learning model for detecting short ChatGPT-generated text," *arXiv preprint*, no. 2301.13852, 2023.
- [27] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is ChatGPT a general-purpose natural language processing task solver?" *arXiv preprint*, vol. 2302.06476, 2023.
- [28] S. Wang, H. Scells, B. Koopman, and G. Zuccon, "Can ChatGPT write a good boolean query for systematic review literature search?" *arXiv preprint*, vol. 2302.03495, 2023.
- [29] K. Canese and S. Weis, "Pubmed: the bibliographic database," *The NCBI handbook*, vol. 2, no. 1, 2013.
- [30] S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner, "Mathematical capabilities of ChatGPT," *arXiv preprint*, vol. 2301.13867, 2023.
- [31] D. Sobania, M. Briesch, C. Hanna, and J. Petke, "An analysis of the automatic bug fixing performance of ChatGPT," *arXiv preprint*, vol. 2301.08653, 2023.
- [32] S. Sanyal, Y. Xu, S. Wang, Z. Yang, R. Pryzant, W. Yu, C. Zhu, and X. Ren, "APOLLO: A simple approach for adaptive pretraining of language models for logical reasoning," *arXiv preprint*, vol. abs/2212.09282, 2022.
- [33] H. Liu, R. Ning, Z. Teng, J. Liu, Q. Zhou, and Y. Zhang, "Evaluating the logical reasoning ability of ChatGPT and GPT-4," *arXiv preprint*, vol. abs/2304.03439, 2023.
- [34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.
- [35] A.-L. Jousselme, G. Pallotta, and J. Locke, "Risk game: Capturing impact of information quality on human belief assessment and decision making," *Int. Journal of Serious Games*, vol. 5, no. 4, pp. 23–44, 2018.
- [36] D. Dubois and H. Prade, *Formal representations of uncertainty*. ISTE, London, UK & Wiley, Hoboken, N.J. USA, 2009, vol. Decision-making - Concepts and Methods, ch. 3, pp. 85–156, invited paper.