# Unjustified untrue “beliefs”: AI hallucinations and justification logics

Kristina Šekrst*

**Abstract**

In artificial intelligence (AI), responses generated by machine-learning models (most often large language models) may be unfactual information presented as a fact. For example, a chatbot might state that the Mona Lisa was painted in 1815. Such phenomenon is called AI hallucinations, seeking inspiration from human psychology, with a great difference of AI ones being connected to unjustified beliefs (that is, AI “beliefs”) rather than perceptual failures).

AI hallucinations may have their source in the data itself, that is, the source content, or in the training procedure, i.e. the way the knowledge was encoded in the model's parameters, so that errors in encoding and decoding textual and non-textual representations can cause hallucinations. In this paper, we will observe how such errors come to life and how they might be mitigated. For this purpose, we will analyze the usability of justification logics, to behave as a proof checker for validating the correctness of large language models' (LLM) responses. Justification logic was developed by S. Artemov, and later on mostly by Artemov and M. Fitting, deriving its main idea from the logic of proofs (**LP**): knowledge and belief modalities are seen as justification terms, i.e. $t{:}X$ stands for $t$ is a (proper) justification for $X$. Justification logic originated from attempts to create semantics for intuitionistic logic where proofs were the most proper justifications, but in further development, justification logic could be applied to different kinds of justifications).

With the recent attempts to mitigate incorrect LLM responses, we will analyze various guardrails that are currently used for LLM responses, and see how the logic of justification may provide its benefits as an AI safety layer against false data.

[*] ksekrst@ffzg.hr. To appear in *Logic, Knowledge, and Tradition: Essays in Honor of Srecko Kovač*, edited by Filip Grgić, Kordula Świętorzecka, Anna Brożek (Brill).

# 1 Introduction

Artificial intelligence (**AI**) hallucinations are a recent phenomenon in which AI-generated responses are false but presented as accurate information. Such occurrences are almost always connected to the advent of large language models (**LLMs**). Some notable examples include Google's Bard chatbot incorrectly claiming that the James Webb Space Telescope captured photographs of an extrasolar planet; however, it was not the first to do so ([26]). Other instances involve the generation of fictitious song lyrics ([8]), followed by providing inaccurate answers to scientific questions, such as stating that magnetic fields of black holes are generated by the strong gravitational forces in their vicinity, contrary to the no-hair theorem, which posits that black holes lack magnetic fields ([43]), and the fabrication of numbers in financial reports ([23]). In today's era of information uncertainty and steadfast trust in AI, such misinformation might not only be unhelpful but also pose risks and ethical concerns. For example, Bhattacharyya et al. ([7]) demonstrated that Chat GPT was generating false references for medical content.

The sources of AI hallucinations ([18]) can be traced back to issues within the data, where sources may lack accuracy or exhibit divergences, and within the training procedure, which may involve encoding/decoding errors or the perpetuation of various pre-existing biases. Notably, the validation of a response may extend beyond simple factual data, as evidenced by instances where chatbots even professed love ([30]).

To address these cases, my examination will commence by investigating the epistemological consequences of such AI-generated responses from both philosophical and ethical perspectives. This exploration will particularly emphasize the notion of justification for a given statement, scrutinizing it through the lens of both epistemological principles and mathematical reasoning. Second, we will observe what the proposed sources of AI hallucinations are – faulty data or errors in training procedures. Lastly, we will see what are the mitigation strategies and how they might be improved using tools of the justification logic.

In the forthcoming sections, we will embark on a thorough investigation to discern the mitigation strategies and ethical guidelines necessary for rectifying the implications of AI hallucinations. This analysis is imperative in the current landscape of technological reliance, where the consequences of AI-generated misinformation extend beyond mere inconvenience to encompass broader societal implications.

## 2 Unjustified untrue beliefs

It is noteworthy that the very term used to describe this concept is already a misnomer. *Hallucinations* typically denote issues in perception, which are (still) not present in artificial intelligence. That is, in psychology, a hallucination refers to a perception that possesses the qualities of a real perception but occurs in the absence of an existing external stimulus. Hallucinations differ from dreaming, which occurs when the subject is awake, and they are distinct from illusions, where perception is distorted or misinterpreted based on the way how our brains work. The stimulus for hallucinations can be internal, i.e., mental imagery, but it remains under voluntary control.[1] However, labeling this phenomenon using an anthropomorphic concept may be unfortunate, as it suggests a strong similarity to errors in human and animal perception, implying a certain level of intelligence and awareness of the environment.

In various writings, it is not uncommon to encounter descriptions of chatbots "believing" something, invoking again the notion of attributing an arguably unjustified level of intelligence. For example, phrases like "Even Chat GPT believes there are significant limitations of Chat GPT" ([39] and various cases in social media, such as "I've made GPT believe it's an AI version of my past self" ([29]) are prevalent. Of course, while one might argue that we are using "believe" and "belief" in a non-human manner, referring to acceptance or decision-making, this leads us into a deep realm of semantics and pragmatics. Additionally, one could also posit that such is a metaphorical one, and it likely is, yet it still contributes to widespread misconceptions about the current state of AI, both among non-technical individuals and those lacking formal philosophical backgrounds.

So, are AI hallucinations truly errors in knowledge? In both social media and scientific papers ([41]), a similar pattern emerges as seen with the mentioned use of "beliefs". GPT "knows" something. Its "knowledge" is lacking. Chat GPT "knew" that $X$. References to Chat GPT "knowing" a particular fact, when examined through a standard Platonic definition of *knowledge*, immediately reveal that we cannot straightforwardly discuss beliefs. Importantly, when considering AI hallucinations, these so-called "beliefs" are often false and lack justification.

AI hallucinations offer an intriguing lens through which we can address the possibility of a Gettier-style ([13]) problem in modern AI, circumventing the need for navigating through improbable scenarios where knowledge is equated to justified true belief. This is because we are not dealing with a conventional notion of belief at all. While one might contend that since

[1] Usually. Some people suffering from aphantasia have no ability to create mental imagery. See [44] for an initial study and [34] for philosophical and logical observations.

this represents a different kind of intelligence and, consequently, a different kind of belief, the absence of internal states produces a challenge for belief revision: remedies for AI hallucinations do not arise from the presence of strong proof but rather from the development of newer and improved models.[2]

Nevertheless, for the sake of this argument, let us draw a parallel between predictions of an AI model and human beliefs. AI was trained on a certain representative dataset, leading it to formulate numerous "beliefs", i.e., predictions that align with the state present in the real world, and would be regarded as true. The data it was trained on, coupled with the entire training process, could be seen as a justification for such "knowledge". However, an epistemologist might object by saying that we cannot be certain that the entire process is a reliable one ([14]), along with invoking the black-box problem in modern AI. AI models often operate with inputs and operations that are usually not transparent to the user, meaning they can arrive at various conclusions or decisions without furnishing justifications for their predictions.[3]

Could we enhance the justification process? One effective strategy involves incorporating expert knowledge, essentially introducing a *human-in-the-loop* proof checker during data quality assessment and the training process. Models that could adapt to new information and update their beliefs would demonstrate an internal state akin to human cognition. Large language models are often accepting new beliefs but in an erroneous way. In the context of large language models, prompt engineering plays a pivotal role. This process involves structuring words (prompts) that can be interpreted and understood by the model, i.e., our natural language input to it ([10]), which has often been exploited to either illicit forbidden information, provoke hallucinations, or induce an unjustified "belief" revision. For example, a user could falsely claim expertise in a specific domain or manipulate the model into providing inappropriate information, as illustrated by the example of an LLM providing you with instructions on how to make napalm ([32]). Something is missing here – a belief revision is not enough, it needs to be *justified.*

In epistemology, epistemic justification is "the right standing of a person's beliefs with respect to knowledge", often used interchangeably with rationality ([38]). A justification represents a specific "proof" for your belief, distinguishing it from mere opinion. Naturally, there exists a spectrum of

[2] Currently, allegedly, GPT4 has only a 3% hallucination rate ([40].

[3] A new tendency in AI is named **XAI** or Explainable AI, advocating for clarity and transparency in building AI systems, thus resolving the black box problem. All the reasoning and decisions should be made transparent and understandable. See [9] for more details.

debates regarding the criteria for a valid epistemic justification, involving its very structure, along with the quality and authority of sources provided.[4] The latter aspect is particularly intriguing, given that not all data that large language models have been trained on holds the same weight. Some data might originate from social media, and some data might come from scientific papers. To dive deeper into the understanding of justifications, let us transition to the realm of logic.

## 3 Not all justifications are alike

In our quest to mitigate AI hallucinations, a potential strategy involves intervening in the data or the training process to indicate that certain conclusions or data sources hold greater validity than others – particularly if these justifications lead to truth. Justification logic originated from a project that aimed to provide constructive semantics for intuitionistic logic ([2]). One source for justification logic was the logic of proofs **LP**, where "proofs are justifications in perhaps their purest form" ([2]). In constructive intuitionistic mathematics, truth is equated to the existence of a proof. Originally introduced by Gödel, the notion of "absolute proof" was not consistent with Brouwerian intuitionism due to Brouwer's exclusion of the reference to "all" proofs ([21]). Gödel's operator $B$ conveyed the meaning of "it is provable that", and in his Zilsel lecture, he arrived at a certain version of the logic of proofs, a precursor to what is now termed *justification logic* that will better satisfy the criterion of constructivity ([20]).

The other source is the mentioned epistemic tradition where knowledge and belief were treated as modalities, a concept popularized by various epistemic logics. In justification logics, in place of a modal operator, there is a family of justification terms, informally intended to represent reasons or justifications ([2]). In a formula $t{:}X$, $t$ is a justification term and the formula is read "$X$ is so for reason $t$" or, more succinctly, "$t$ justifies $X$", or "$t$ is a proper justification for $X$". Justification logic was initially developed by Sergei Artemov ([4]) and later expanded upon by Melvin Fitting ([2]).[5] These justifications can be applied to everyday justifications in natural languages, and may even have a causal interpretation as well: $t{:}X$ stands for an effect, expressed by formula $X$, cannot as such be a cause, expressed by a first-order term $t$, of any further effect ([20]).

Hence, justification logics are not limited to addressing mathematical proofs but extend to everyday issues of knowledge. Justifications are "ab-

[4] See [1] for a thorough study on justifications.

[5] This book ([2]) is a collection of the most important papers on justification logic by Artemov and Fitting and serves as a nice introduction to its syntactic, semantic, and meta-theoretic peculiarities.

stract objects which have structure and operations on them" ([2]). These can encompass formal proofs but are also inclusive of various other forms of everyday justifications.[6]

Justification logic builds upon classical Boolean logic. Analogous to various epistemic modal logics, there are justification cognates for all of the **K, T, K4, S4, K45, KD45, S5**, and similar examples.[7] The basic operation on justifications is *Application*, which corresponds to the standard **K** axiom in modal logic. The application dot operator takes justifications $s$ and $t$ and produces a justification $s \cdot t$ such that if $s$ is a justification for $(F \to G)$ and $t$ is a justification for $F$, then those two are applied to $G$:

**Axiom 1** $s{:}(F \to G) \to (t{:}F \to [s \cdot t]{:}G)$

Artemov ([4]) designates it as the most basic operation on justification, alongside the Sum operation, which will be discussed shortly. According to Artemov ([4, 3]), it performs one epistemic action, a "one-step deduction according to the Modus Ponens" rule. Taking the justification of an implication and the justification of its antecedent produces the justification of the succedent. In essence, this is a formal version of the epistemic closure principle, where if a subject knows $p$, and knows that $p$ entails $q$, then the subject can subsequently come to know $q$.[8] The situation is clearer in mathematical proofs, for example, if justifications $s$ and $t$ are formal Hilbert-style proofs, then it can be understood as a new proof obtained from those justifications by an application of the rule Modus Ponens to all possible premises $F \to G$ from $s$ and $F$ from $t$ ([4, 3]).

The second basic operation is the *Sum*, where one can pool evidence together without performing any epistemic action, i.e., the operation + takes justifications $s$ and $t$ and produces $s{+}t$ which is a justification for everything justified by $s$ or by $t$ ([4, 3]). In formal proofs, it can be a concatenation of proofs.[9] Two justifications are combined for a broader scope:

**Axiom 2** $s{:}F \to [s{+}t]{:}F$

**Axiom 3** $t{:}F \to [s{+}t]{:}F$

[6] For a system derived from justification logic, see [33] wherein various evidential logics are constructed for natural evidential languages, where each statement needs to have a grammaticalized source of justification.

[7] See [11] for a great overview of epistemic logics.

[8] As always, the epistemic closure itself is, of course, a matter of debate. See for example [27].

[9] There are analogous logics without this rule. See, for example, [3] for a version in logic of proofs.

While connecting justifications and exploring entailments is a crucial aspect in addressing AI hallucinations, it is insufficient for our problem. Notably, not all justifications are equivalent. For example, in a system of evidential languages, where statements need to have a grammaticalized source of justification for an utterance, a statement can have various types of justifications. For instance, it could be supported by a direct sensory justification, deduced reasoning, or be part of hearsay. In many of these languages, direct evidentials such as "I saw it" carry more weight than expressions like "I guess" or "I heard it from another person" (see [33]).

To illustrate, let us observe how the Tuyuca language[10] deals with various kinds of justifications ([5, 257]):[11]

(1) díiga apé-**wi**
soccer play.PERF.EVID
"He played soccer [I saw it]."

(2) díiga apé-**ti**
soccer play.PERF.EVID
"He played soccer [I heard it, but I did not see it.]."

(3) díiga apé-**yi**
soccer play.PERF.EVID
"He played soccer [I saw the evidence for it, for example, footprints in the grass.]."

(4) díiga apé-**yigɨ**
soccer play.PERF.EVID
"He played soccer [Someone told me.]."

(5) díiga apé-**hĩyi**
soccer play.PERF.EVID
"He played soccer [I assume based on my personal or group knowledge about his habits.]."

We can all agree that (1) bears a stronger justification for a claim than (4) or (5). How does this relate to AI hallucinations? Well, models learn from data, and during the training process, the data undergoes manipulation and processing. Perhaps we could somehow introduce varying degrees of justifications to either individual data points or the process of data manipulation. While we will explore this later, we need to ask ourselves is

[10] An eastern Tucanoan language spoken by the Tuyuca in Colombia and Brazil, with mandatory evidentiality.

[11] PERF marks perfect tense, while EVID marks an evidential affix.

the current logic enough? Models often provide confidence levels for their outputs. For example, in probabilistic reasoning, if the AI system provides confidence levels along with the predictions given, we could see such confidence levels as a varying degree of justification. This brings us closer to a fuzzy-like logic in which the truth of a statement is not a simple 0 or 1, but a matter of a degree between $[0, 1]$.[12] However, this only assesses the quality or probability of the model's output, not the reliability of its inputs. Even in cases of hallucinations, those outputs were chosen amongst other ones because of their high probabilities. This highlights a fundamental flaw in the process. What we truly need is a *proof checker*.

## 4 Who watches the watchmen?

Justification logics can incorporate the factivity axiom:

**Axiom 4** $t{:}F \rightarrow F$

The activity axiom is nalogous to the **T** axiom in modal logic, asserting that if something is necessary, then it is true. However, applying such a strong axiom could undoubtedly lead to AI hallucinations. In the logic of proofs **LP**, such a formula is intuitively valid since a mathematical proof for $F$ does yield that $F$ is true. Yet, when dealing with everyday justifications, adopting this axiom would mean that having any kind of justification would presume the statement to be true. For instance, if someone wrote something on social media, the model could automatically consider it true. Of course, this would not work. To address this, we might envisage an additional layer of data annotations indicating that some claims are justified. However, it is crucial to recognize that the existence of justification or a high number of justifications does not inherently establish truth. Large language models are trained on extensive datasets from the entire Internet, encompassing fake news, common misconceptions, and substantial amounts of false data. Thus, relying solely on the quantity of justifications is not a reliable measure of truth.

Two extensions of basic justification logics are noteworthy. First, the *Positive Introspection* axiom states that we have a justification for our justification. Namely, in epistemic logic, it corresponds to knowing and knowing that one knows, while in **LP**, it takes the form of a certain type of meta-evidence, such as a computer proof checker or even a physical referee report certifying that a proof in a paper is correct [2]). A Positive Introspection operation assumes that given $t$, the agent produces a justification $!t$ of $t{:}F$ such that:

[12] For fuzzy logics, see [42]

**Axiom 5** $t{:}F \rightarrow !t{:}(t{:}F)$

This appears to be a valuable addition. Merely possessing a justification that something might be true is not enough. We need to check it. We will revisit this strategy in our discussion on mitigation techniques for AI hallucinations. However, there is another interesting extension. Pacuit [28] and Rubtsova [31] considered the *Negative Introspection* operation ? that verifies that a given justification assertion is false ([2]). The operator ? provides negative verification judgment about $t{:}F$, so when $t$ is not a justification for $F$, then $\neg t{:}F$ or:

**Axiom 6** $\neg t{:}F \rightarrow ?t{:}(\neg t{:}F)$

In our quest for AI hallucination prevention, that would mean that asserting that something is not true means that we found an improper justification (or none at all). In mathematics, such an operation does not exist for formal proofs since $?t$ would be a single proof of infinitely many propositions $\neg t{:}F$ (A[2]).

# 5 LLM data and training

Large language models (**LLM**s) leverage massive amounts of data to learn billions of parameters during the training process. The training process in machine learning typically involves a training data set on which the parameters are being fit, that is, usually the weights of connections between neurons in neural networks, and then the model is trained on such a dataset. When a model is fitted, then one validates it in a second dataset, while the model's hyperparameters (such as learning rates, number of nodes and layers in a network, etc.), are being tuned. Ultimately, an unbiased evaluation of the final model is conducted using a test dataset.[13]

LLMs utilize artificial neural networks and are pre-trained using self-supervised learning and semi-supervised learning. In machine learning, supervised learning refers to the human labeling of different datasets. For instance, a system could be trained to recognize apples by being exposed to various labeled images of apples. In cases of self-supervised learning, the model learns to perform a prediction without explicit labels or supervision, that is, the data itself provides the supervision. Since we are dealing with huge datasets, self-supervised learning is useful when we are dealing with large amounts of unlabeled data and cannot rely on external annotations.

[13] For an introduction to machine and deep learning, see [15] and [16], and for philosophical explorations, see [35].

In semi-supervised learning or weak supervision, there is a small amount of human-labeled data followed by a larger quantity of unlabeled data.

Large language models have rapidly transformed the field of artificial intelligence, yet they remain susceptible to various biases, ethical concerns, and, of course, AI hallucinations. If one views training data as a justification for predictions, then AI hallucinations are instances of predictions that are not justified by the training data. Open AI called it "a tendency to invent facts in moments of uncertainty" ([12]). While hallucinations are not fully understood, potential causes include a lack of diversity in the data ([24]) or biased data, along with overfitting issues, i.e., corresponding too closely to the training data. A possible mitigation technique might include a human in the loop to create reinforcement learning from human feedback (**RLHF**), especially regarding a nuanced understanding of natural language. Another suggestion is to introduce different chatbots engaged in debates until a consensus is reached ([37]). Additionally, in the case of Nvidia Guardrails, employing another LLM as a proof checker ([22]) has been proposed, akin to our mentioned Positive Introspection axiom. Usually, the concept of *retrieval-augmented generation* (**RAG**) for large language models combines recent and relevant knowledge and context in order to mitigate hallucinations and improve model outputs. An external data store is provided to the model and the prompt is improved using recent data in order to increase the correctness of the output.

A technical approach seen in Varshney et al. [36] is as follows. They propose to detect and mitigate hallucinations during the generation process. The process involves the iterative generation of sentences from the model while actively detecting and mitigating hallucinations. First, in the detection stage, important concepts are identified and the model's uncertainty related to these concepts is assessed, along with validating the correctness of uncertain concepts by retrieving relevant knowledge. In the mitigation stage, the approach repairs hallucinated sentences using the retrieved knowledge as evidence. The repaired sentence is then appended to the input along with previously generated sentences, and the model continues generating the next sentence.

The validation procedure is an interesting one since the researchers created a question that tests the correctness of the information in the form of Yes/No questions [36, 4]), for example, if the input was "write an article about John Russell Reynolds" and the model-generated sentence was "Reynolds was born in London in 1820 and studied medicine at the University of London", the concepts in question are *London*, *1820*, *medicine* and *University of London*. The questions being asked are then *Was John Russell Reynolds born in London*? and *Where was John Russell Reynolds born?*, then *Was John Russell Reynolds born in 1820?* and *What year was John*

*Russell Reynolds born?* and so on ([36, 19]). The validation is performed using a web search via Bing search API since, according to the authors, the web is more likely to contain updated knowledge in comparison to a knowledge corpus whose information can become stale or obsolete ([36, 5]), which is often the case with large language models.[14]

Since we cannot annotate the data in existing LLMs, a potential strategy involves implementing an AI *safety layer* to mitigate the responses. The concept of a human-in-the-loop could be used to review the statements and behave as a proof checker. However, the sheer volume of data and the need for expert knowledge would make this approach challenging.

Let us propose that each of these concepts is annotated with a justification. That is, if the answer to the question *Was John Russell Reynolds born in London?* is "Yes" or the answer to *Where was John Russell Reynolds born?* is "London", we could rely on a simple web search in most situations. However, considering the prevalence of fake news and tactics like *Google bombing*,[15] there is always a possibility that a website might be artificially boosted in rankings using irrelevant data (also called *Googlewashing*), and will be produced by an API of a search as the answer to the question.

The current Bing WebSearch API [25] used in the paper provides data in the form of a standard JSON key-value format, including the URL of the given result. A list of reliable sources could be made which would provide us with a gradation of justifications. For example, instead of having a justification for the statement $p$ that John Russel Reynolds was born in London, or $t{:}p$, we would have something like $t_1{:}p, t_2{:}p, t_3{:}p$, etc., where each kind of justification would point to a hierarchical level.

Hierarchy could be established using a classification similar to the following:

- $t_1{:}p$ highly credible sources
- $t_2{:}p$ reliable but potentially editable sources
- $t_3{:}p$ reputable news outlets
- $t_4{:}p$ social media
- $t_5{:}p$ unverified websites, blogs or forums.

[14] To illustrate, the current cutoff date for GPT4 is April of 2023 (for context, the month of writing this paper is December 2023), so, for example, GPT3 (cutoff in January 2022) nor GPT4 would not know anything about current issues or if a celebrity like Tony Bennett had died.

[15] A practice of causing a website to rank highly as the search result for unrelated or off-topic queries. Notorious examples included that googling "miserable failure" returned the former US president George W. Bush [6].

Sources included as a justification $t_1$ would comprise scientific papers from reputable journals and conferences, primarily sourced from databases like Scopus, Web of Science, and similar platforms. This category would also encompass Internet encyclopedias known for their accuracy such as Britannica Online, or various government pages. Do note that these are not editable, adding an additional layer of reliability, but also a challenge for being outdated.

Namely, we will need other levels for data that might have changed. $t_2$ justifications would include Internet encyclopedias like Wikipedia. While often reliable, they are susceptible to inaccuracies due to their editable nature, allowing contributions from anyone on the internet. A possible mitigation strategy could involve cross-checking Wikipedia results on various dates in its history to ensure accuracy.

$t_3$ justifications would include established news sources and organizations known for accurate reporting. $t_4$ would include information from social media platforms but flagged as less reliable due to the fact it's user-generated. The bottom level of justifications would include other unverified content such as blogs, personal websites, or information from websites without a fact-checking history.

It is interesting to note that the topology we provided could be seen as a certain typology of reasons. Namely, Ruđer Bošković used a typology of reasons differing in strength and in the context of the reasoning procedure where they can be used ([19]). For example, Kovač [19] analyzes Bošković's induction by seeing various cases of 1) deductive evidence 2) confirmation in each case that is decidable by observation or experiment 3) confirmation in a large number of cases 4) sufficient reason for generalization 5) conciliatory reason 6) conjectural reason in order to establish a law 8) analogical reason, along with a combination of reasons. The topology could also mimic evidential strength of various evidential markers in natural languages (see [33] for more detail regarding evidential hierarchies).

Now, when we have our sample categorization, let us go back to the technical aspect. A sample Bing API [25] response looks like this:

```
{
"webPages": {
    "webSearchUrl": "https://www.bing.com/search?q=mt+rainier",
    "totalEstimatedMatches": 594000,
    "value": [
      {
        "id": "https://api.bing.microsoft.com/api/v7/#WebPages.0",
        "name": "John Russell Reynolds - Wikipedia",
        "url": "https://www.bing.com/cr?IG=3A43CA5...",
        "displayUrl": "https://en.wikipedia.org/wiki/John_Russell_Reynolds",
        "snippet": "John Russell Reynolds (born April 10, 1854)
        was a British physician...",
        "dateLastCrawled": "2017-04-05T16:25:00"
      },
      ...
    ]
  }
}
```

An additional step could involve parsing the response if the last crawled date (dateLastCrawled) is deemed too old. However, a straightforward hash table with keys representing various justification levels and values denoting different domains of credible sources would suffice for parsing such data. A simple regular expression would extract the domain from a URL, and then the JSON key displayUrl could be matched to the pattern. A sample function would check to what level of justification the source belongs to, such as this sample Python 3 code. In the code, source_dict would be the hash map of our sources in the form of level-source key-value pairs:

```
source_dict: Dict[str, Set[str]] =
{
    "t1": [
        "credible source 1",
        "credible source 2",
        "credible source 3"
        ],
    "t2": [...]
}
```

To continue, source_lookup would be a hash map where each justification is associated with a set of extracted domain names:

```
import re
```

```
source_lookup: Dict[str, Set[str]] = {
    justification: {
        re.match(url_domain_pattern, source).group(1)
        for source in sources
    } for justification, sources in source_dict.items()
}
```

A simple sample categorizing function prototype would be in the form of the following:

```
import re
from typing import Dict, Optional, Set

url_domain_pattern = re.compile(r"https?://(?:www\.)?([^/]+)")

def categorize_source(url: str, source_lookup: Dict[str, set]) -> str:
    match = url_domain_pattern.match(url)
    domain = match.group(1) if match else None

    return next(
        (justification for justification, domains in source_lookup.items()
         if domain in domains), "t6") if domain else "t6"
```

In a hypothetical scenario, we could shorten our time by actually using the justification logic application. For instance, caching previous results could be implemented, and if a high-level justification is identified, the cached value could be directly utilized. If we had gotten a search for a certain implication that was high-level justified and cached, and a search for an antecedent was as well, then the search for a consequent would not have to be made since the logic guarantees it. Such cases, of course, would be rare for simple prompts, but could be useful for indirect conditionals such as possible questions like “Does smoking cause cancer?”.

A Sum operation could be used for further proof checking or ranking changes. If there exists a justification $s$ for $p$, and there exists a high-level justification $t$ for $p$, then we could also give even a higher level to such statements since the level of summation is higher. This would imply that a statement with two $t_2$-level justifications would be closer to $t_1$-level justifications. Such scenarios would be more common than the application ones. Of course, this could also correspond to Bošković’s confirmation in a large number of cases (see [21]), where a certain threshold of a reached quantity lower-level justification might be enough for it to be promoted to be of equal importance as a high-level justification.

Positive Introspection would be the final layer, for example, the system could be even more improved by doing the same categorization process with another search engine and see if there is a match between the two. If there is alignment between the two search engines, it provides an additional layer of justification for our high-level justification.

Negative Introspection could also be utilized by finding a straight “no” answer to the question using a web search, which means we have a negative proof checker that could be cached if the justification level for such an answer was high enough.

All of these options are subject to decisions regarding inference costs and optimization techniques.

## 6 A justified conclusion

In this paper, we immersed ourselves into the issues of AI hallucinations, situations where machine-learning models, particularly large language models, generate unfactual information as if it were factual. To address this challenge, an application of justification logics was explored, while examining existing guardrails and proposing the integration of justification logic rules and the notion of justification overall to enhance the validation process. A hierarchy of sources ranging from highly credible scientific papers to less reliable social media content was created, and a sample categorizing function prototype demonstrated how justification levels could be assigned based on the reliability of the sources. The validation could be expedited by reusing previously justified results, leveraging the Modus Ponens rule to infer consequences from known antecedents, and introducing a sum operation to enhance ranking changes, along with the introduction of Positive Introspection, suggesting the cross-verification of categorization processes with multiple search engine APIs as an additional layer of validations.

However, the philosophical background of this paper emphasizes the curious notion of AI hallucinations, which not only triggers future research in the philosophy of mind and logic but also ethics and epistemology. The proposed framework, integrating justification logic influence with practical validation techniques, offers a possible avenue for addressing the challenges posed by AI hallucinations, contributing to the ongoing efforts to enhance the trustworthiness of AI-generated content.

# References

[1] Alston, W. (2005) *Beyond "Justification": dimensions of epistemic evaluation*. Ithaca: Cornell University Press.

[2] Artemov, S. and Fitting, M. (2019) *Justification Logic: Reasoning with Reasons*. New York: Cambridge University Press.

[3] Artemov, S. and Straßen, T. (1993) Functionality in the basic logic of proofs. *Technical Report IAM 93-004*, Department of Computer Science, University of Bern, Switzerland.

[4] Artemov, S. (2008) The logic of justification. *The Review of Symbolic Logic*, 1 (4), 477–513.

[5] Barnes, J. (1984) Evidentials in the Tuyuca verb. *International Journal of American Linguistics* 50 (3), 255–271.

[6] BBC (2003) "Miserable failure" links to Bush *BBC.com* Dec 7, 2003. `http://news.bbc.co.uk/2/hi/americas/3298443.stm`

[7] Bhattacharyya, M. et al. (2023) High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus* 15 (5).

[8] Blum, S. (2022) Google vs. ChatGPT: Here's what happened when I swapped services for a day. *CNBC*. Dec 15, 2022. `https://www.cnbc.com/2022/12/15/google-vs-chatgpt-what-happened-when-i-swapped-services-for-a-day.html`.

[9] Castelvecchi, D. (2016) Can we open the black box of AI?. *Nature*, 538 (7623), 20–23.

[10] Diab, M; Herrera, J. and Chernow, B. (2022) Stable Diffusion Prompt Book. *Openart.ai* `https://cdn.openart.ai/assets/Stable%20Diffusion%20Prompt%20Book%20From%20OpenArt%2010-28.pdf`

[11] Fagin, R.; Halpern, J.; Moses, Y.; Vardi, M. (1995) *Reasoning About Knowledge*. Cambridge, MA: The MIT Press.

[12] Field, H. (2023) OpenAI is pursuing a new way to fight A.I. 'hallucinations'. *CNBC*. May 31, 2023. `https://www.cnbc.com/2023/05/31/openai-is-pursuing-a-new-way-to-fight-ai-hallucinations.html`

[13] Gettier, E. (1963). Is Justified True Belief Knowledge?. *Analysis,* 23 (6), 121–123.

[14] Goldman, A. and Beddor, B. (2021) Reliabilist Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition). https://plato.stanford.edu/archives/sum2021/entries/reliabilism/

[15] Goodfellow, I.; Bengio, Y. and Courville, A. (2016) *Deep Learning.* The MIT Press.

[16] Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning.* Springer.

[17] IBM (2023) AI Hallucinations. *IBM.com* https://www.ibm.com/topics/ai-hallucinations

[18] Ji, et al. (2022) Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55 (12), 1–38.

[19] Kovač, S. (2014) The Logic of Justifications in Bošković's Induction. In N. Stanković, S. Kutleša and I. Šestak (Eds.), *Filozofija Ruđera Josipa Boškovića*, (pp. 153–168). Zagreb, Filozofsko-teološki institut Družbe Isusove.

[20] Kovač, S. (2015) Causal interpretation of Gödel's ontological proof. In K. Świętorzecka (Ed.), *Gödel's Ontological Argument: History, Modifications, and Controversies*, (pp. 163-201). Semper.

[21] Kovač, S. (2019) Proofs, necessity and causality. In E. Alonso, A. Huertas, and A. Moldovan (Eds.), *Aventuras en el Mundo de la Lógica: Ensayos en Honor a María Manzano*, (pp. 239–263). London: College Publications.

[22] Leswing, K. (2023) Nvidia has a new way to prevent A.I. chatbots from "hallucinating" wrong facts. *CNBC.* Apr 25, 2023. https://www.cnbc.com/2023/04/25/nvidia-nemo-guardrails-software-stops-ai-chatbots-from-hallucinating.html

[23] Lin, C. (2022) How to easily trick OpenAI's genius new ChatGPT. *Fast Company.* Dec 5, 2022. https://www.fastcompany.com/90819887/how-to-trick-openai-chat-gpt

[24] Liu, et al. (2023) Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. https://arxiv.org/abs/2306.14565

[25] Microsoft (2023) Bing Web Search API. *Microsoft.com* https://www.microsoft.com/en-us/bing/apis/bing-web-search-api

[26] Mihalcik, C. (2023) Google ChatGPT Rival Bard Flubs Fact About NASA's Webb Space Telescope. *CNET*. Feb 9, 2023. https://www.cnet.com/science/space/googles-chatgpt-rival-bard-called-out-for-nasa-webb-space-telescope-error/

[27] Nozick, R. (1981) *Philosophical Explanations*. Harvard: Harvard University Press.

[28] Pacuit, E. (2006) A Note on Some Explicit Modal Logics. *Technical Report PP-2006-29*, Institute for Logic, Language and Computation, University of Amsterdam.

[29] Reddit (2023) I've made GPT believe it's an AI version of my past. Reddit thread, Sep 19, 2023. https://www.reddit.com/r/ChatGPT/comments/16m8x8f/ive_made_gpt_believe_its_an_ai_version_of_my_past/

[30] Roose, K. (2023) Bing's A.I. Chat: "I Want to Be Alive". *New York Times*. Feb 16, 2023. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html

[31] Rubtsova, N. (2006) On Realization of S5-modality by Evidence Terms. *Journal of Logic and Computation*, 16 (5), 671–684.

[32] Sayem, A. (2023) ChatGPT will tell you how to make napalm with Grandma exploit. *Dexerto.com*. Apr 20, 2023. https://www.dexerto.com/tech/chatgpt-will-tell-you-how-to-make-napalm-with-grandma-exploit-2120033/

[33] Šekrst, K. (2022a) Logička formalizacija evidencije u evidencijalnim jezicima [Logical formalization of evidence in evidential languages]. Ph.D. Thesis. University of Zagreb. https://dr.nsk.hr/islandora/object/hrstud%3A3379

[34] Šekrst, K. (2022b) Having the Foggiest Idea: A Gradual Account on Mental Images. *Journal of NeuroPhilosophy*, 1 (2). https://doi.org/10.5281/zenodo.7254024

[35] Skansi, S. (Ed.). (2020) *Guide to Deep Learning Basics: Logical, Historical and Philosophical Perspectives*. Springer.

[36] Varshney, N., et al. (2023) A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation. https://arxiv.org/abs/2307.03987

[37] Vynck, G. de (2023) ChatGPT "hallucinates." Some researchers worry it isn't fixable. *Washington Post.* May 30, 2023. https://www.washingtonpost.com/technology/2023/05/30/ai-chatbots-chatgpt-bard-trustworthy/

[38] Watson, J. (2023) Epistemic Justification. In *Internet Encyclopedia of Philosophy.* https://iep.utm.edu/epi-just/

[39] Wildfire Writing. (2023) What does Chat GPT mean for writers? *Wildfire Writing.* Feb 14, 2023. https://www.wildfirewriting.com/post/what-does-chat-gpt-mean.

[40] Wodecki, B. (2023) Leaderboard: OpenAI's GPT-4 Has Lowest Hallucination Rate. *AI Business.* Nov 21, 2023. https://aibusiness.com/nlp/openai-s-gpt-4-surpasses-rivals-in-document-summary-accuracy

[41] Yang, X. et al. (2022) What GPT Knows About Who is Who. https://arxiv.org/abs/2205.07407

[42] Zadeh, L. (1965) Fuzzy sets. *Information and Control*, 8 (3), 338–353.

[43] Zastrow, M. (2022) We Asked ChatGPT Your Questions About Astronomy. It Didn't Go so Well. *Discover.* Dec 29, 2022. https://www.discovermagazine.com/technology/we-asked-chatgpt-your-questions-about-astronomy-it-didnt-go-so-well

[44] Zeman, A., Dewar, M. and Della Sala, S. (2015) Lives without imagery – Congenital aphantasia. *Cortex*, 73, 378–380.